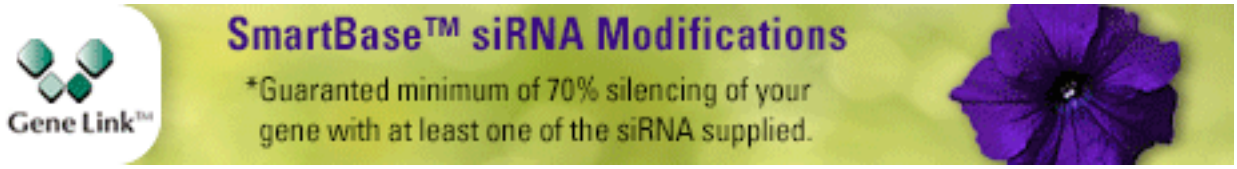
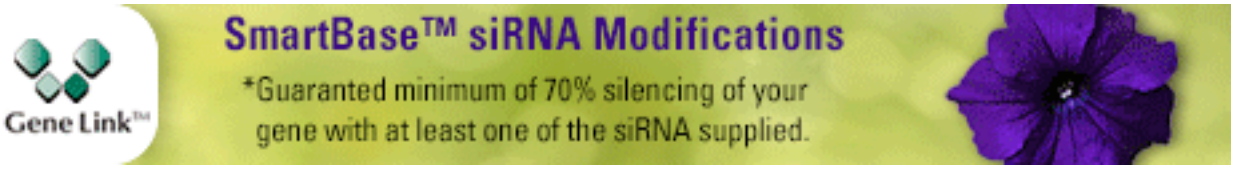
Comparison of *D. melanogaster* and *C. elegans* developmental stages, tissues, and cells by modENCODE RNA-seq data

Jingyi Jessica Li, Haiyan Huang, Peter J. Bickel, et al.

Genome Res. 2014 24: 1086-1101
Access the most recent version at doi:10.1101/gr.170100.113





Comparison of *D. melanogaster* and *C. elegans* developmental stages, tissues, and cells by modENCODE RNA-seq data

Jingyi Jessica Li,[1,3] Haiyan Huang,[1,4] Peter J. Bickel,[1,4] and Steven E. Brenner[2,4]

[1]*Department of Statistics, University of California, Berkeley, California 94720, USA;* [2]*Department of Plant and Microbial Biology, University of California, Berkeley, California 94720, USA*

We report a statistical study to discover transcriptome similarity of developmental stages from *D. melanogaster* and *C. elegans* using modENCODE RNA-seq data. We focus on "stage-associated genes" that capture specific transcriptional activities in each stage and use them to map pairwise stages within and between the two species by a hypergeometric test. Within each species, temporally adjacent stages exhibit high transcriptome similarity, as expected. Additionally, fly female adults and worm adults are mapped with fly and worm embryos, respectively, due to maternal gene expression. Between fly and worm, an unexpected strong collinearity is observed in the time course from early embryos to late larvae. Moreover, a second parallel pattern is found between fly prepupae through adults and worm late embryos through adults, consistent with the second large wave of cell proliferation and differentiation in the fly life cycle. The results indicate a partially duplicated developmental program in fly. Our results constitute the first comprehensive comparison between *D. melanogaster* and *C. elegans* developmental time courses and provide new insights into similarities in their development . We use an analogous approach to compare tissues and cells from fly and worm. Findings include strong transcriptome similarity of fly cell lines, clustering of fly adult tissues by origin regardless of sex and age, and clustering of worm tissues and dissected cells by developmental stage. Gene ontology analysis supports our results and gives a detailed functional annotation of different stages, tissues and cells. Finally, we show that standard correlation analyses could not effectively detect the mappings found by our method.

[Supplemental material is available for this article.]

Drosophila melanogaster and *Caenorhabditis elegans* are model systems for studying molecular, cellular, and developmental processes in animals (Wolpert 2011). As morphologically different and evolutionarily distant organisms separated by as much as 600 million years in evolution (Adoutte et al. 2000; Weigmann et al. 2003), *D. melanogaster* and *C. elegans* have striking differences in cell differentiation and whole-organism developmental biology (Lettre and Hengartner 2006; Lesch and Page 2012). Besides the obvious differences in their morphological changes and developmental timelines (Fig. 1A,B; Supplemental Fig. S1), additional differences exist in their development, including, for example: (1) *C. elegans* has an alternative developmental path—dauer-interrupted development—a state of developmental arrest that does not exist in the life cycle of *D. melanogaster*; (2) adult *D. melanogaster* has males and females of equal proportions, whereas adult *C. elegans* has 99.5% hermaphrodites and only 0.05% males; (3) *D. melanogaster* has a pupal stage, in which the great majority of larval differentiated tissues are histolyzed, and the adult is formed from previously undifferentiated tissues; whereas *C. elegans* goes through only one major cycle from undifferentiated to differentiated tissues; (4) in contrast to *D. melanogaster*, *C. elegans* has a highly invariant embryonic lineage, which gives rise to specific cell fates; and (5) the number of nuclei in syncytial *D. melanogaster* embryos exceeds the number of somatic cells in adult *C. elegans*. Despite these differences, many individual conserved mechanisms have been observed in *D. melanogaster* and *C. elegans*, such as asymmetric cell division (Betschinger and Knoblich 2004), cell migration, and axon pathfinding (Montell 1999); and of course these species contain many similar histological cell types common to all animals. Indeed, the conservation of embryonic development in animal species has been a unifying concept since von Baer's observations in the 19th Century (Kalinka and Tomancak 2012), and the conservation of developmental genes between animals has long been studied in evolutionary developmental biology, e.g., the *Hox* genes (Pearson et al. 2005). However, we know of no genome-wide analyses to systematically characterize the conservation in gene expression during the development and cell differentiation of *D. melanogaster* and *C. elegans*.

Genome-wide mRNA expression profiling surveys have shown that gene expression changes accompany morphological changes in the development of both *D. melanogaster* and *C. elegans* (e.g., Jiang et al. 2001; Kim et al. 2001; Arbeitman et al. 2002; Stolc et al. 2004; Kalinka et al. 2010). Such studies have also observed similarities in gene expression between some *D. melanogaster* early and late developmental stages (Arbeitman et al. 2002), between some cell lines from *D. melanogaster* female adults and early embryos (Cherbas et al. 2011), and between *C. elegans* dissected cells and their corresponding developmental stages (Spencer et al.

[3]Present address: Department of Statistics, University of California, Los Angeles, California 90095, USA
[4]Corresponding authors
E-mail brenner@compbio.berkeley.edu
E-mail bickel@stat.berkeley.edu
E-mail hhuang@stat.berkeley.edu

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.170100.113. Freely available online through the *Genome Research* Open Access option.

24:1086–1101 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/14; www.genome.org

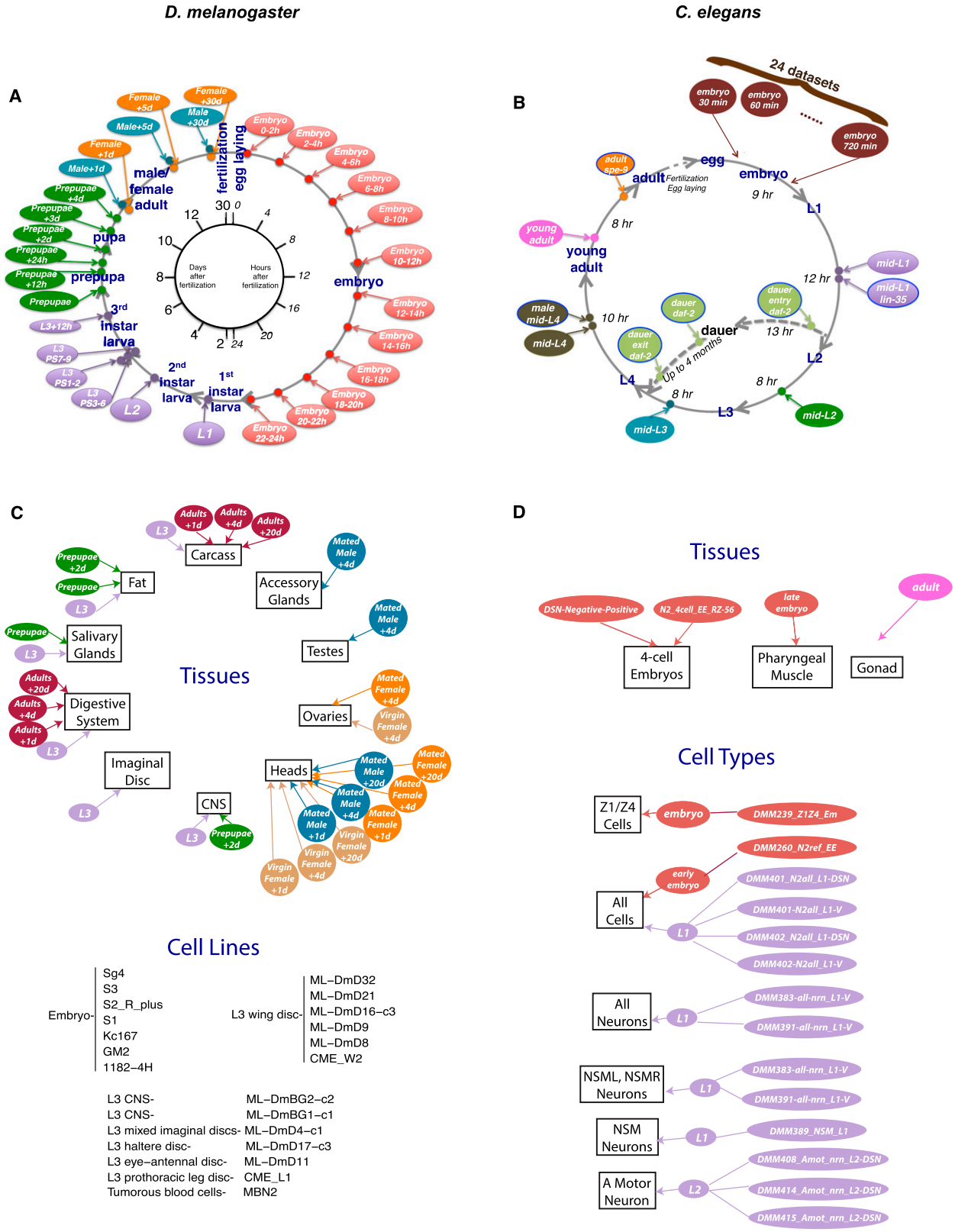


Figure 1. Life cycles and modENCODE RNA-seq data sets of *D. melanogaster* and *C. elegans* (Gerstein et al. 2014). (*A*) modENCODE RNA-seq data sets of 30 different *D. melanogaster* developmental stages. (*B*) modENCODE RNA-seq data sets of 35 different *C. elegans* developmental stages. (*C*) modENCODE RNA-seq data sets of 29 tissues and 19 cell lines in *D. melanogaster*. (*D*) modENCODE RNA-seq data sets of four tissues and 14 dissected cells in *C. elegans*. For detailed information on the stage, tissue, and cell labels, please refer to Supplemental Table S2.

2011). A comprehensive comparison of multiple developmental stages within and between the two species has not been conducted nor have the stages been compared with tissues or cells across species.

The Model Organism ENCyclopedia Of DNA Elements (modENCODE) Project (Celniker et al. 2009) provides an unprecedented resource for studying genome-wide gene expression patterns in multiple *D. melanogaster* (fly) and *C. elegans* (worm) developmental stages, tissues, and cells (including fly cultured cell lines and worm dissected cells). High-throughput RNA sequencing (RNA-seq) data from 131 biological samples (stages, tissues, and cells) with more than 11 billion total aligned reads in the two species are available (Gerstein et al. 2010, 2014; The modENCODE Consortium et al. 2010). The fly time-course data include 30 developmental stages, from embryos, L1–L3 larvae, pupae, to male and female adults (Fig. 1A). The worm time-course data contain 35 developmental stages, including embryonic, L1–L4 larval, young adult, adult, and dauer stages (Fig. 1B). The tissue and cell data (Fig. 1C,D) include 29 fly tissues of 10 types (carcass, fat, salivary glands, digestive system, imaginal discs, CNS, heads, ovaries, testes, accessory glands), 19 fly cultured cell lines, four worm tissues, and 14 worm dissected cells. Each of those samples was characterized by RNA-seq with a minimum of 9 million reads and a median of 73 million reads. For labels and more information on the biological samples, please refer to Supplemental Table S2.

In this paper, we used these data to compare developmental stages, tissues, and cells of *D. melanogaster* and *C. elegans* based on genome-wide protein-coding gene expression. Our comparison approach centers on using orthologous genes to link the two species. We identify "associated genes" to capture the transcriptional characteristics of different biological samples (i.e., developmental stages, tissues, and cells). Within fly or worm, we find the similarity of ("map") two biological samples by looking for a significant overlap in their associated genes. Between fly and worm, we map two biological samples if they exhibit shared transcriptional characteristics, i.e., a significant proportion of their associated genes are orthologous. Using this approach, we performed an extensive mapping of developmental stages, tissues, and cells within and between fly and worm. The within-species mapping results are consistent with previous findings (Arbeitman et al. 2002; Cherbas et al. 2011; Spencer et al. 2011), thus supporting the validity of our approach; they also provide new information on the similarity and differences of various stages, tissues, and cells within fly and worm. More importantly, the between-species mapping results reveal previously unknown correspondence in transcription between the life cycles of fly and worm, and also show novel relationships among fly and worm stages, tissues, and cells. Our results provide—for the first time to our knowledge—a comprehensive map between *D. melanogaster* and *C. elegans* developmental stages, tissues, and cells, indicating that conservation exists in their development and tissue/cell differentiation.

Results

Identifying stage-/tissue-/cell-associated genes to map stages/tissues/cells based on transcriptome characteristics

To find if there is any transcriptome similarity among the developmental stages, tissues, and cells of *D. melanogaster* and *C. elegans*, we started by identifying genes whose expression captures the transcriptome characteristics of a particular biological sample (i.e., a stage, tissue, or cell). This is motivated by the fact that genes with constant expression across all biological samples, e.g., a few housekeeping genes, provide little information to differentiate the transcriptomes of various biological samples. Specifically, for every developmental stage, we selected its "stage-associated genes" as the genes relatively highly expressed at that stage compared to other stages throughout development; for every tissue/cell, we similarly selected its "tissue-/cell-associated genes" as the genes highly expressed in that tissue/cell relative to other tissues/cells. We define the associated genes of a sample by the following criterion: the genes that have FPKM (fragments per kilobase of transcript per million mapped reads) ≥ 1.0 and Z-score (normalized FPKM across samples) > 1.5 in the sample (see Methods for details). This criterion guarantees that in the given sample, the selected associated genes are expressed at levels distinguishable from background noise and are also more highly expressed than in some other samples. We use these associated genes as a basis to compare the developmental stages, tissues, and cells within and between *D. melanogaster* and *C. elegans*. For between-species comparison, we focus on using orthologous genes (i.e., genes in different species but originated from a single gene of their last common ancestor) to link the two species, and we restrict the fly-/worm-associated genes to those that have orthologs in worm/fly.

In this study, we use *D. melanogaster* and *C. elegans* gene annotations from Ensembl (version 66) (Flicek et al. 2012) and orthologous genes from modENCODE (Table 1; Supplemental Table S1; Wu et al. 2014). Expression of protein-coding genes at different developmental stages or in different tissues/cells was estimated from modENCODE RNA-seq data (Fig. 1; Supplemental Table S2) by using Cufflinks (Trapnell et al. 2010). We identified associated genes of different stages, tissues, and cells from the gene expression estimates (see Methods for detailed identification criteria).

The number of associated genes for a given sample ranges from ~300 to ~4500, whereas the number of associated genes with orthologs ranges from ~100 to ~1600 (Supplemental Fig. S2). The stages and tissues/cells with higher transcriptional activities, such as early embryonic stages and genital glands, generally have more associated genes. Supplemental Figure S2 also illustrates how many genes are associated with more than one biological sample. For most stages, their associated genes are also associated with about three neighboring stages. This phenomenon agrees with the facts that gene expression changes continuously during development (Holter et al. 2000) and that the organisms could not have been perfectly synchronized. There are also biological samples, such as worm "L4 male" and fly "testes mated male + 4d," which have a large proportion of uniquely associated genes (Supplemental Fig. S2B,C,F,G). This agrees with the apparent morphological differences of L4 male worms and fly testis tissues from all the other biological samples. We note that worm embryos at 90, 120, 150, and 180 min are not of equally good quality as the other worm embryo samples (L Hillier, pers. comm.). This may explain why we observe a spike in the number of associated genes of worm embryos at 150 min.

We compare two biological samples by statistically checking the dependence of their associated genes: If the two samples are within the same species, we test the significance of the number of their common associated genes; if the two samples are from different species, we test the significance of the number of orthologous gene pairs in their associated genes. We call the two samples "mapped" if their associated genes have significant dependence (Bonferroni-corrected P-value $< 10^{-3}$ from a hypergeometric test, in which the null hypothesis is that the two samples have in-

Table 1. Summary of *D. melanogaster* and *C. elegans* protein-coding genes and orthologs

				Number of fly genes	Number of worm genes	Number of pairs
Protein-coding genes with orthologs	Ortholog pair type (fly-worm)	1–1		3131	3131	3131
		1–many	1–2	310	620	620
			1–3	79	237	237
			1–4	37	148	148
			1–≥5	53	465	465
		Many–1	2–1	618	309	618
			3–1	234	78	234
			4–1	76	19	76
			≥5–1	262	32	262
		Many–many	2–2	132	132	264
			2–≥3	76	154	308
			≥3–2	136	60	272
			≥3–≥3	323	354	4768
	Total			5467	5739	11,403
	All protein-coding genes[a]			13,781	20,389	

Orthologs are from modENCODE prediction of fly-worm orthologs (http://compbio.mit.edu/modencode/orthologs/modencode-orths-2012-01-30/ensembl-v65/modencode.merged.orth.txt.gz).
[a]*D. melanogaster* genome assembly: BDGP 5.64 (Ensembl assembly 66); *C. elegans* genome assembly: WS 220 (Ensembl assembly 66).

dependent associated genes; in other words, the two samples are unrelated) (see Methods for details), implying great similarity in their transcriptome characteristics. Figure 2 illustrates our comparison strategy.

Within-species comparison of *D. melanogaster* developmental stages and tissues/cells

We first applied this strategy to comparing developmental stages, tissues and cultured cell lines within *D. melanogaster*. Figure 3A summarizes the results for 30 fly developmental stages. First, as expected, a diagonal pattern is observed, i.e., adjacent developmental stages are mapped to each other. Second, we found a few off-diagonal mappings: early embryos/female adults (i.e., early embryos are mapped to female adults), middle embryos/larvae, and late embryos/pupae. The observed mapping between the earliest embryonic stage (i.e., embryo 0–2 h) and female adult stages (i.e., female adult 5–30 d) is consistent with an independent study (Cherbas et al. 2011). The other two mappings, one observed between middle embryos (i.e., embryo 10–16 h) and larvae (i.e., L1, L2) and the other between late embryos (embryo 14–18 h) and pupae (prepupae + 2–3 d), both agree with previous microarray profiling analysis (Arbeitman et al. 2002). It is known that in the fly development, most larval cells die in metamorphosis, and pupal tissues are generated from imaginal discs, which are progenitor cells allocated in embryogenesis and remain quiescent during embryonic and larval stages. Hence, it is not surprising that we observe that the pupal stages are more similar to late embryos than larvae. Those reasonable stage-mapping results support the validity of our approach.

To determine whether the mapping of early embryos to female adults is a result of maternal gene expression in oocytes, we further compared the 30 developmental stages with three gene categories defined by Lott et al. (2011). In that paper, the researchers used strain-specific time series of *D. melanogaster* gene expression at eight embryonic time points to classify 9003 fly genes into three categories: 5598 maternal genes (whose transcripts are maternally deposited), 2210 zygotic genes (whose transcripts are zygotically expressed), and 1195 maternal + zygotic genes (whose transcripts are both maternally deposited and zygotically expressed). Similar to our stage comparison strategy, we performed a hypergeometric test on the overlap of stage-associated genes with the genes in each category. Figure 3B shows that the associated genes of early embryos and female adults are significantly enriched in the maternal gene category, and to a lesser extent in the maternal + zygotic gene category. On the other hand, the associated genes of the other developmental stages are mostly enriched in the zygotic gene category. This result confirms that the mapping of early embryos to female adults is attributable to the expression of maternal and maternal + zygotic genes in these two developmental time periods.

Figure 3C summarizes the comparison results of 48 fly tissues and cultured cell lines (reordered by hierarchical clustering) (see Methods for details), where we observed a clear separation of tissues and cell lines. Remarkably, the cell lines

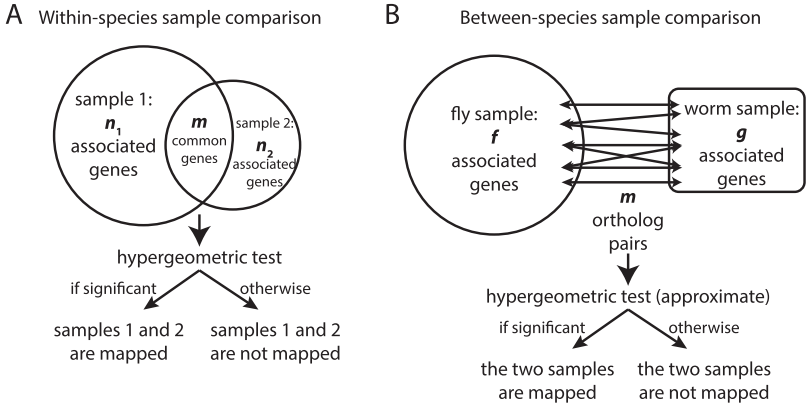


Figure 2. Approaches for comparing transcriptomes of samples. (*A*) Approach for comparing two biological samples within a species. A hypergeometric test is used to test whether the overlap in their associated genes is significant. (*B*) Approach for comparing two biological samples between two species. An approximate hypergeometric test is used to test the significance of the number of orthologous gene pairs in their associated genes.

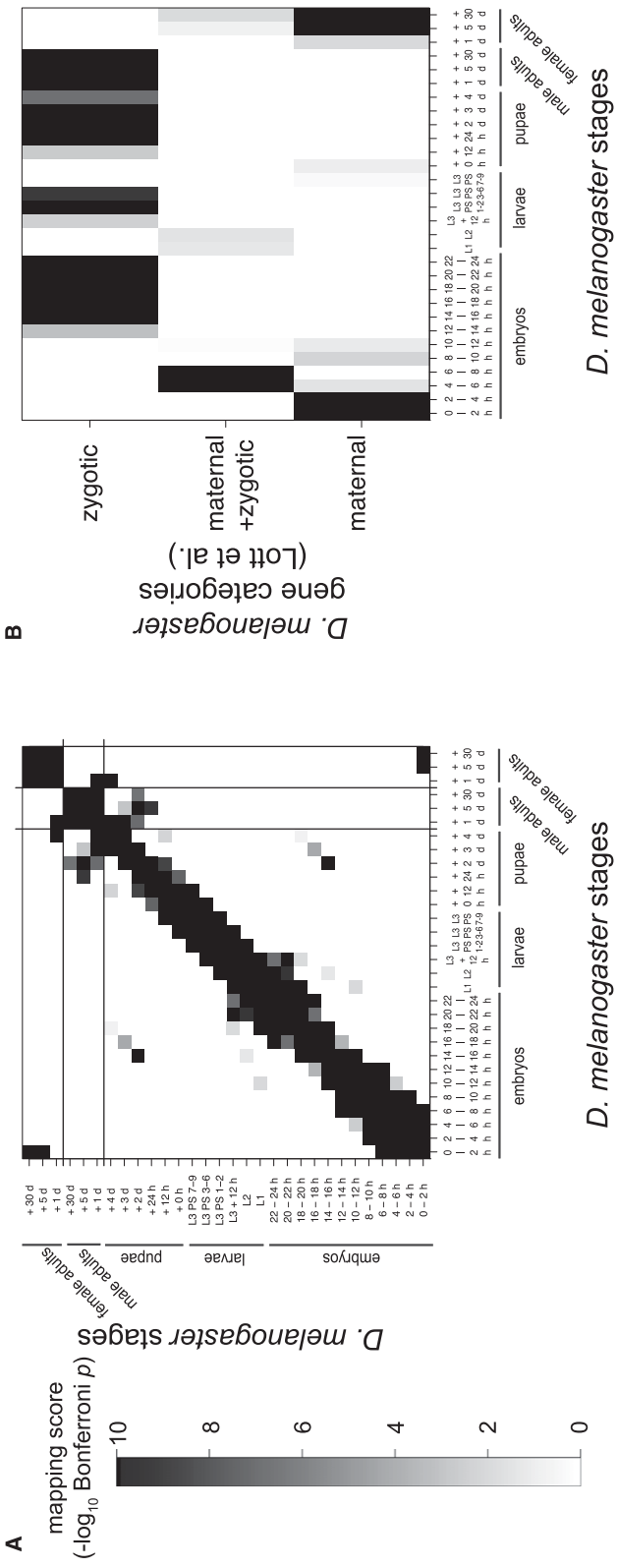


Figure 3. (Continued on next page)

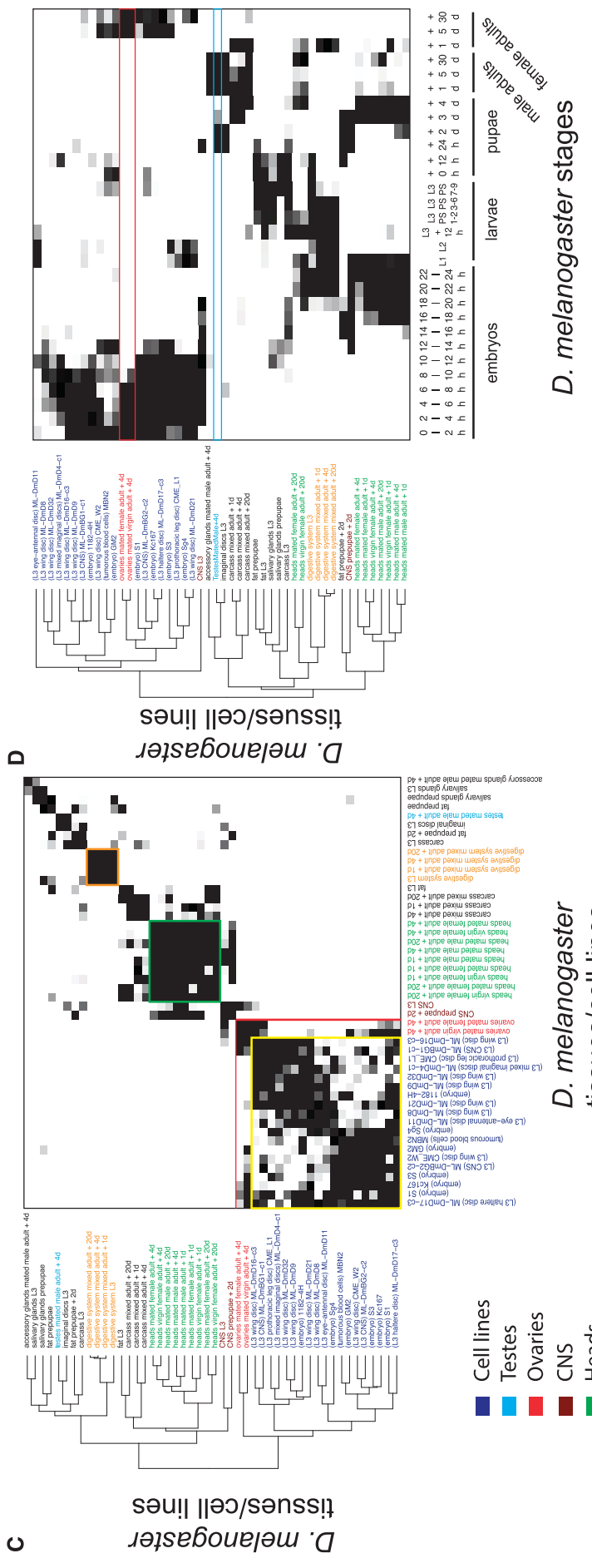


Figure 3. Comparison results of different developmental stages, tissues, and cell lines within *D. melanogaster*. More significant mapping scores are shown in darker color, corresponding to the scale showing $-\log_{10}$ transformed Bonferroni corrected *P*-values, which are calculated from the hypergeometric test in Figure 2A. (*A*) Stage comparison result. All are from mixed organisms with the exception that adults are separated into male and female that have the same three developmental time points (1, 5, and 30 d after eclosion). (*B*) Comparison of developmental stages with the three gene categories (maternal, maternal/zygotic, and zygotic) defined in Lott et al. (2011). (*C*) Tissue/cell line comparison result. The grouping of cell lines (yellow box), the mapping of cell lines and ovary tissues (red box), the grouping of head tissues (green box), and the grouping of digestive system tissues (orange box) are highlighted. (*D*) Comparison of developmental stages with tissues/cell lines, in which the red box marks the mapping of ovary tissues to early embryonic and female adult stages, and the cyan box marks the mapping of testes tissues to pupa and male adult stages. Hierarchical clustering was applied to order the tissues/cell lines in C and *D*. Tissues from similar organs and cell lines are marked with colors. For detailed information on the stage, tissue, and cell labels, please refer to Supplemental Table S2. Mapping score as in *A* and *B*.

form a strong grouping with each other (yellow box) rather than with their originating tissues, indicating that cultured cell lines share certain transcriptome characteristics not found in tissues. This is a reasonable result because cell lines are more highly proliferative than tissues and thus have higher expression of genes involved in growth and cell cycling, a phenomenon that has been well established in mammalian cell lines and tumors (Whitfield et al. 2002). Mapping is also observed between ovary tissues and cell lines (between the yellow box and the red lines), as a result of the high expression of maternal expressed genes (Supplemental Fig. S3). It is also consistent with the reported similarity between cell lines and early embryos (Cherbas et al. 2011) and our observed mapping of early embryos to female adults due to maternal gene expression in ovaries (Fig. 3A,B). In addition, Figure 3C shows that different fly adult head tissues (heads of mated male, mated female, and virgin female adults+ 1, 4, and 20 d) are mapped to each other (green box), and so are the digestive system tissues of mixed adults at different time points (adults + 1, 4, and 20 d, orange box). These results revealed substantial similarity in gene expression among fly adult tissues of the same type irrespective of differences in sex and age.

We next compared the 48 fly tissues and cell lines with the 30 developmental stages. From the results in Figure 3D, we first observed a clear mapping of the ovary tissues to both early embryonic and female adult stages (in red box), which again confirms that the maternal genes highly expressed in oocytes lead to the observed mapping of early embryos to female adults in the stage comparison (Fig. 3A,B). We next observed a mapping of the testes tissues to a few mixed pupal stages and the three male adult stages (in cyan box). We also found an interesting block formed between all the cell lines and early embryonic stages (Fig. 3D, top left) and another block between most of the cell lines and female adult stages (top right). Combined with our previously observed similarities between cell lines and ovaries and between early embryonic and female adult stages, these results depict a transcriptome similarity module comprised of fly cell lines, ovary tissues, and early embryonic stages as well as female adults that harbor eggs in the ovaries.

To determine common biological processes between the mapped *D. melanogaster* stages/tissues/cells lines, we found the biological process (BP) gene ontology (GO) terms that are significantly enriched in the associated genes of every fly stage, tissue, and cell line (see Methods for detail). Supplemental Figures S4–S6 show the enrichment patterns of these significantly enriched BP GO terms across all stages or all tissues/cell lines. We observe that the mapped stages and the mapped tissues/cell lines have common enriched GO terms, which provide functional explanation of the mapping result. For example, GO terms related to cell division and germ cell development are highly enriched in both 0–2 h embryos and female adults (Supplemental Fig. S4A). In ovaries, heads, and the digestive system, different isolates of each tissue share GO terms related to cell division, neuronal functions, and metabolic processes, respectively (in Supplemental Fig. S4B, see the red, green, and orange boxes, respectively). We did further functional analysis to provide a detailed annotation of the biological functions enriched in each fly developmental stage and tissue/cell line (Supplemental Fig. S7; Supplemental Materials). Many functions are enriched in expected stages, such as cell division in 0–2 h embryos and adult females; regulation of gene expression in 8–10 h embryos; locomotion and behavior in 22–24 h embryos; histolysis and catabolism in prepupae; and mating in adult males. In addition, less obvious processes are also found to be stage enriched. For example, we are not aware of any previous evidence that immune response genes would be most highly expressed in prepupae or that carbohydrate metabolism and energy production would be particularly prevalent in adult males. We also found enriched BP GO terms in common associated genes between mapped cell lines, between cell lines and early embryos, and between cell lines and female adults (Supplemental Table S5C; see Methods for detail). Many of the enriched GO terms are related to cell cycling and growth, confirming that those samples are highly proliferative. To allow further analysis, we have provided a complete list of stage/tissue/cell-associated genes (Supplemental Table S3) for all fly developmental stages and tissues/cell lines.

Within-species comparison of *C. elegans* developmental stages and tissues/cells

We next applied the same strategy to comparing developmental stages, tissues, and cells within *C. elegans*. Figure 4A summarizes the mapping results of 35 worm stages, from which we inferred three interesting patterns. First, as expected, we generally observed mappings between adjacent worm stages in the time course. Second, mappings are found between embryos (0–120 min incubation after egg harvest) and adult stages. Since worms are ~99.5% hermaphrodites that produce all their sperm in the L4 stage and then switch over to producing oocytes in adults (Nayak et al. 2005), the observed mapping of early embryos to adults is likely due to maternal gene expression in worm oocytes, analogous to what was seen in the fly. Third, dauer stages are mapped to the two L1 larval stages. Dauer stages are an alternative developmental pathway that starts after stage L1 and exits into stage L4. Thus, our mapping of dauers to L1 agrees with their temporal proximity in the *C. elegans* life cycle (Fig. 1B; Supplemental Fig. S1B).

Figure 4B summarizes the comparison results of 18 worm tissues and dissected cells. After hierarchical clustering, we found that tissues and cells from similar origins show strong groupings: Cells dissected from L1-stage worms (in red) are mapped together; embryonic tissues/cells (in green) are mapped to each other. These findings are consistent with earlier reports that worm dissected cells have a similar gene expression as their corresponding tissues (Spencer et al. 2011). We also observed that 4-cell embryos are mapped to adult gonad tissues, further supporting the idea that our mapping of early embryos to adults (Fig. 4A) is likely attributable to gene expression in gonad tissues and perduring maternal mRNA in early embryos.

We next compared the 18 worm tissues and dissected cells with the 35 developmental stages. In the results (Fig. 4C), tissues and dissected cells are generally mapped in understandable ways to the corresponding developmental stages. First, we observed mappings between embryonic tissues/cells (in green) and early embryonic stages. Second, 4-cell embryos/adult gonad tissues are mapped to early embryonic and adult stages (in red box). Third, we found mappings between dissected cells from L1-stage worms (mixed cells [in red] and neurons [in blue]) and late embryonic to larval developmental stages. These observed gene expression similarities between worm tissues/cells and corresponding developmental stages are consistent with previous findings by principal component analysis on microarray data (Spencer et al. 2011). However, in fly, although we observed such similarity between some fly tissues (e.g., fat, carcass, and salivary glands) and their corresponding developmental stages, other fly tissues (e.g., heads and digestive systems) and cell lines are not mapped to their corresponding stages (Fig. 3D). Possible reasons include differences in the nature of developmental programs of fly and worm, differences

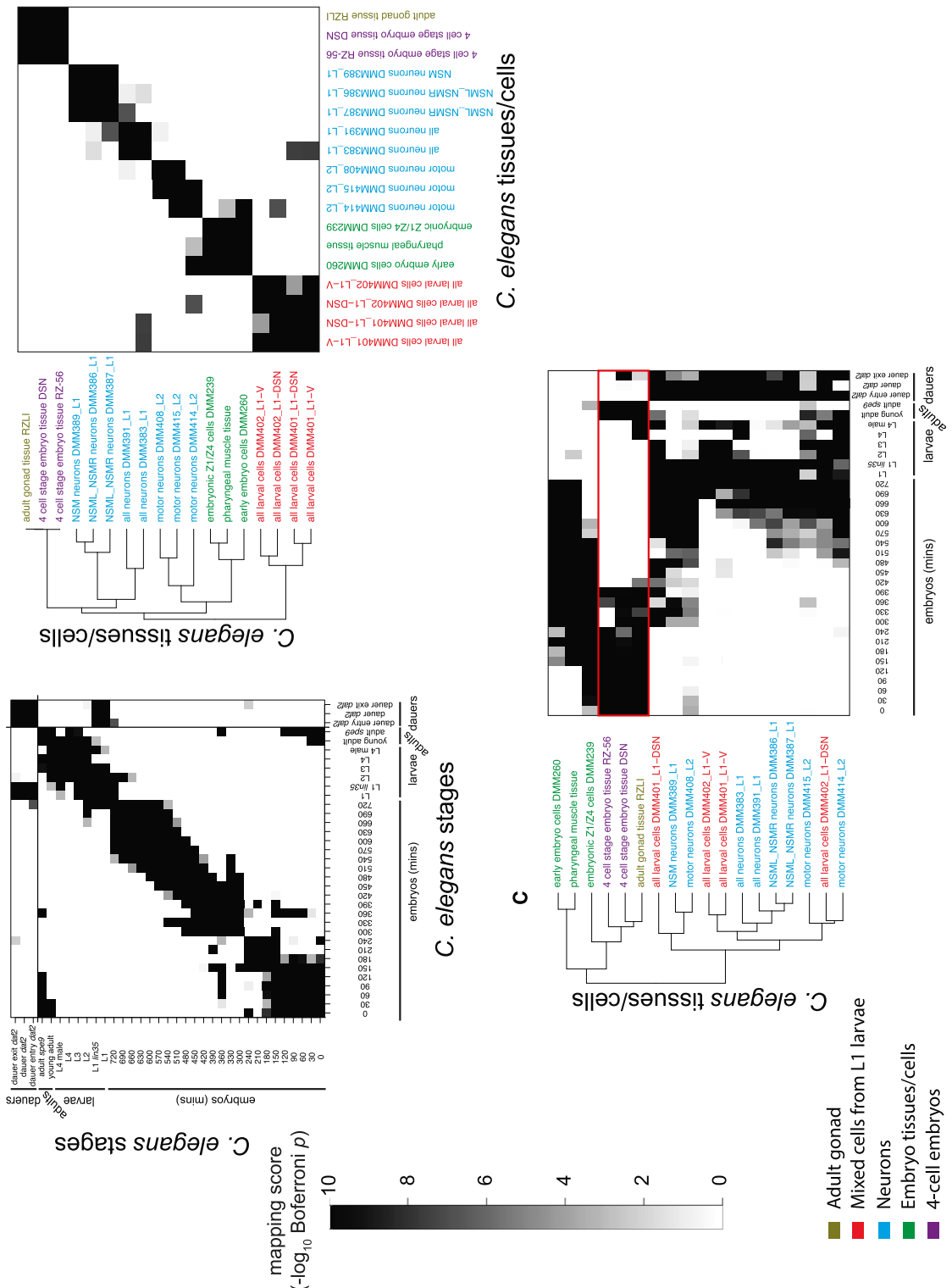


Figure 4. Comparison results of different developmental stages, tissues and dissected cells within *C. elegans*. Mapping scores shown in the heatmaps are $-\log_{10}$ transformed Bonferroni corrected *P*-values, which are calculated from the hypergeometric test in Figure 2A. (*A*) Stage comparison result, in which the three dauer stages are put on the *right* because they form an alternative developmental path in the worm life cycle. (*B*) Tissue/cell comparison result. (*C*) Comparison of developmental stages with tissues/cells, in which the red box marks the mappings of 4-cell embryos and adult gonad tissues to both embryonic and adult stages. Hierarchical clustering was applied to order the tissues/cells in *B* and *C*. Tissues and cells with similar origins are marked with the same color. For detailed information on the stage/tissue/cell labels, please refer to Supplemental Table S2.

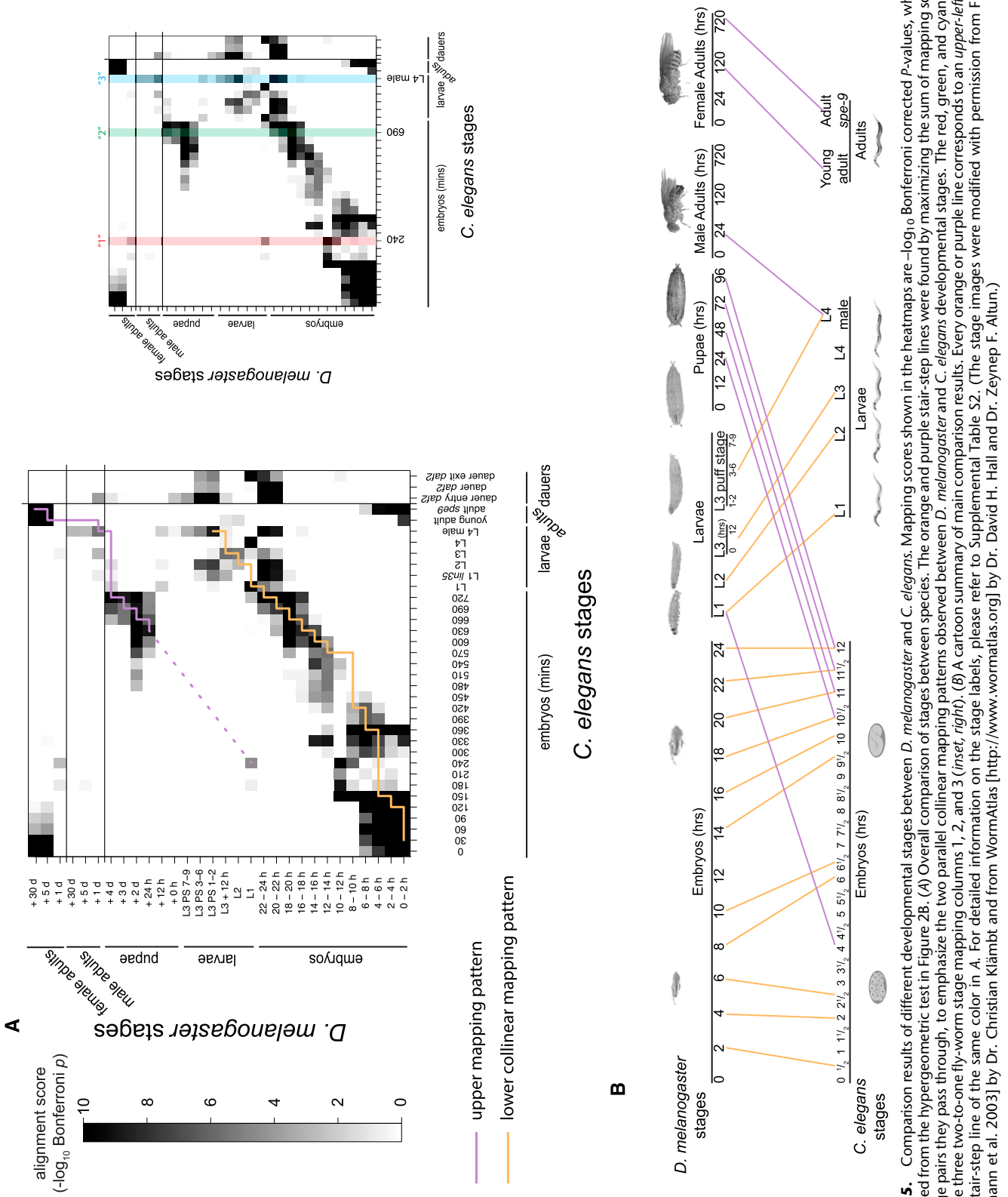


Figure 5. Comparison results of different developmental stages between *D. melanogaster* and *C. elegans*. Mapping scores shown in the heatmaps are $-\log_{10}$ Bonferroni corrected *P*-values, which are calculated from the hypergeometric test in Figure 2B. (*A*) Overall comparison of stages between species. The orange and purple stair-step lines were found by maximizing the sum of mapping scores of the stage pairs they pass through, to emphasize the two parallel collinear mapping patterns observed between *D. melanogaster* and *C. elegans* developmental stages. The red, green, and cyan shades mark the three two-to-one fly-worm stage mapping columns 1, 2, and 3 (*inset, right*). (*B*) A cartoon summary of main comparison results. Every orange or purple line corresponds to an *upper-left* corner in the stair-step line of the same color in *A*. For detailed information on the stage labels, please refer to Supplemental Table S2. (The stage images were modified with permission from FlyMove [Weigmann et al. 2003] by Dr. Christian Klämbt and from WormAtlas [http://www.wormatlas.org] by Dr. David H. Hall and Dr. Zeynep F. Altun.)

in the organisms' scale and anatomy, and the fact that the fly cell lines are immortalized, whereas the fly and worm tissues and worm dissected cells are not.

To determine the biological processes prominent in each sample, we calculated the enrichment of BP GO terms in either all worm developmental stages or all worm tissues/cells. The enrichment patterns of the BP GO terms that are highly enriched in at least one stage or one tissue/cell are summarized in Supplemental Figures S8–S10. Like the analysis in *Drosophila*, these results, combined with our further functional analysis in Supplemental Figure S11, provide biological bases to our mapping results as well as a functional annotation of genes associated with different developmental stages and tissues/cell lines. For example, early embryonic and young adult stages are both enriched with GO terms related to growth and reproduction (Supplemental Fig. S8A), and so are 4-cell stage embryos and adult gonad tissue (Supplemental Fig. S8B); biological processes related to dauer entry and response to food are among the 10 most enriched functions in dauers and are also well enriched in late stage embryos and L1 larvae; embryos "EE_50-240" are enriched for genes annotated for gonad development; and embryos "EE_50-690" are enriched in neurogenesis genes (Supplemental Fig. S11A). A complete list of associated genes at every worm developmental stage or tissue/cell is given in Supplemental Table S3. For more information, see Supplemental Material.

Between-species comparison of *D. melanogaster* and *C. elegans* developmental stages and tissues/cells

We then applied our strategy to comparing the developmental stages, tissues, and cells between the two species. As the first study on possible global correspondence between the life cycles of *D. melanogaster* and *C. elegans*, we compared their developmental stages on the basis of shared orthologs in their stage-associated genes. Figure 5A shows a striking stage mapping result that we analyze in detail here (see Supplemental Fig. S12 and Supplemental Material for more details) and also report in brief in the accompanying integrative modENCODE transcriptome paper (Gerstein et al. 2014). First, a collinear pattern is observed between fly early embryonic through larval stages and worm early embryonic through larval stages. Second, another more fragmentary parallel pattern is found, formed by four sets of fly-worm stage pairs: fly L1 larvae/worm middle embryos, fly prepupae/worm late embryos, fly male adults/worm L4 male, and fly female adults/worm adults. Figure 5B presents a cartoon that summarizes the stage mapping results, showing the two parallel patterns as a division of the fly life cycle into two parts: The first part (from fly early embryos to larvae) is aligned with the complete worm life cycle except for the worm adults (orange lines), and the second part (from fly prepupae to adults) is aligned with the worm life cycle except for the worm early embryos (purple lines).

As a direct result from the two parallel patterns, it is interesting to find that several worm stages are mapped to two blocks of fly stages, e.g., worm embryos at 240 and 690 mins and male L4 larvae (columns 1, 2, and 3 in Fig. 5A, inset). We call the mapping of a worm stage to an earlier block of fly stages as "early mapping," and the mapping of the worm stage to a later block of fly stages as "late mapping." The presence of early and late mappings is consistent with the fact that flies undergo two rounds of large-scale proliferation of undifferentiated tissue followed by differentiation: once during embryonic stages and again during larval and pupal stages (Fig. 1A; Supplemental Fig. S1A). By detailed analysis on the genes that lead to the three example columns 1, 2, and 3, we found evidence suggesting that fly pupal development may have evolved in part by use of duplicated genes involved in late embryonic development. First, we compared the fly stages in the three example columns 1, 2, and 3 with the within-fly stage mapping results in Figure 3A. We found that the two blocks of fly stages in column 1 (middle embryos and L1 larvae) are moderately mapped to each other within fly and so are the two fly stage blocks in column 2 (middle to late embryos and pupae). However, the two fly stage blocks in column 3 (late embryos to prepupae and male adults) are not mapped within fly, suggesting that different fly genes may be expressed in the two blocks. In other words, the off-diagonal stage mappings in Figure 3A cannot fully explain all the two-to-one fly-worm stage mapping patterns in Figure 5A. More detailed analysis on the fly-worm orthologs leading to such patterns reveals that the "early mapping" and "late mapping" in columns 2 and 3 are largely attributable to many-to-one fly-worm orthologs. In particular, in many cases a single gene in the common ancestor to worms and flies has given rise to only one daughter gene in worm but has duplicated in the fly lineage, resulting in one paralog expressed in early fly development and another paralog expressed at a later stage. Overall, this suggests duplication and subfunctionalization of significant portions of the fly developmental program (see Supplemental Material).

In addition to the two parallel stage mapping patterns, we also observed mappings between fly early embryos and worm adults, and between fly female adults and worm early embryos (Fig. 5A). These results, coupled with the mapping between fly female adults and worm adults, indicate strong orthology in the maternal oocyte genes of the two species. Moreover, worm dauer stages are mapped with fly late embryos, larvae, and male adults.

To discover the biological processes involved in the above fly-worm stage mapping, for every worm stage, we picked either the fly stage it is most strongly mapped to (if the worm stage is not involved in the two parallel patterns) or the most strongly mapped early and late fly stages (if the worm stage is involved in the parallel patterns). A complete summary of the fly and orthologous worm genes that are enriched in each stage pair and their GO terms is available in Supplemental Table S7A. Most stage pairs have distinctive biological functions, with exceptions in the early and late mappings in each of columns 1, 2, and 3 (Fig. 5A), where many gene functions are related (see Supplemental Material). For column 1, translation and RNA processing genes are common in the early and late mappings; for column 2, ion transport, neuronal function, and behavior; and for column 3, ion transport and protein phosphorylation (Supplemental Table S7C). Interestingly, the genes common in the early and late mappings of column 1 include many ribosomal proteins, twofold reductions in the expression of which have been long known to significantly lower cell division rates in a cell autonomous manner (Ashburner et al. 2005), and disruption of ribosomal proteins can lead to the "minute" class of mutations in fly (Marygold et al. 2007). Thus our observations that these and other general metabolic genes show stage-specific increases in expression imply that even the regulation of relatively general housekeeping genes could play focused roles in important developmental processes.

To understand the similarity of tissues/cells and developmental stages between the two species, we use the same between-species stage comparison approach to compare (1) worm tissues/cells with fly stages (Fig. 6A), and (2) fly tissues/cells to worm stages (Fig. 6B). Figure 6A shows that worm 4-cell embryos and adult gonad tissues are mapped to both fly early embryonic and female adult stages (red box). Figure 6B shows that fly gonad tissues

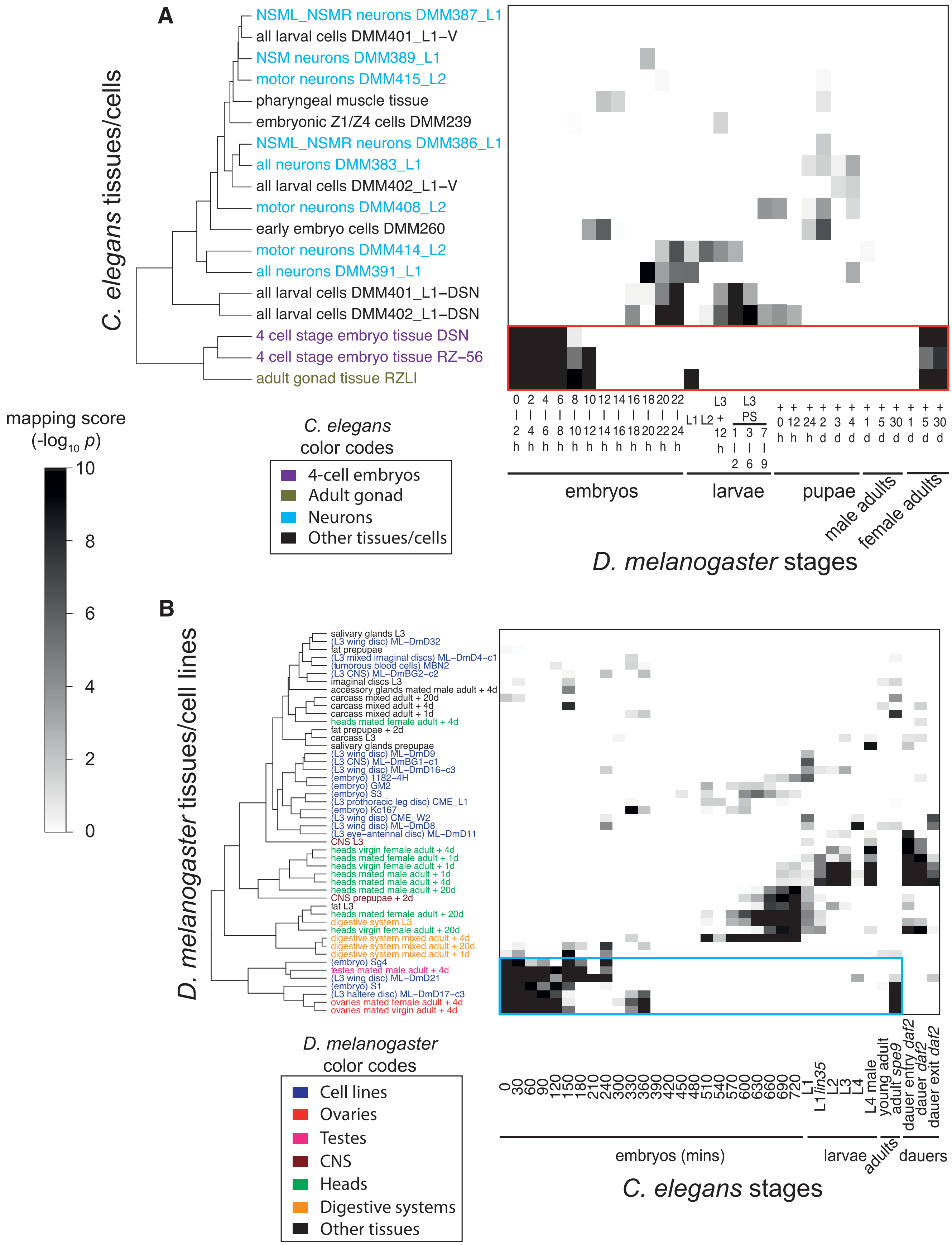


Figure 6. (Continued on next page)

(ovaries and testes) and several cell lines are mapped well to worm early embryonic and adult stages (blue box). These two findings are again results of maternal oocyte gene expression, implying again strong orthology between maternal oocyte genes and their expression in both species.

Finally, we compared tissues and cells between fly and worm. The comparison results in Figure 6C show two interesting patterns. First, most worm neuron tissues are mapped to fly heads in adults and CNS tissues in larvae and pupae (green box), indicating strong orthology of genes with neural functions in both species. Second, worm 4-cell embryos and adult gonad tissues have a clear mapping to many fly cell lines and adult ovaries and testes (yellow box). Digestive tissues in fly are mapped to two worm larval mixed cell samples with duplex-specific nuclease (DSN) treatment, but are surprisingly not mapped to the other two larval mixed cell samples without DSN treatment. In addition, several tissues such as salivary glands and carcass in fly, which may not have natural correlates in worm, are not mapped to any tissues or cells in worm.

Stage comparison based on gene ontology

From the perspective of functional studies, we performed additional comparison of fly and worm developmental stages based on gene ontology (GO) instead of orthology to link the two species. Specifically, we mapped two stages if their associated genes exhibit significant overlap in their corresponding GO terms (specifically, leaf nodes of the biological process gene ontology tree) (see Methods for detail). Supplemental Figure S13, A and B, shows that the within-species comparison results exhibit similar but much

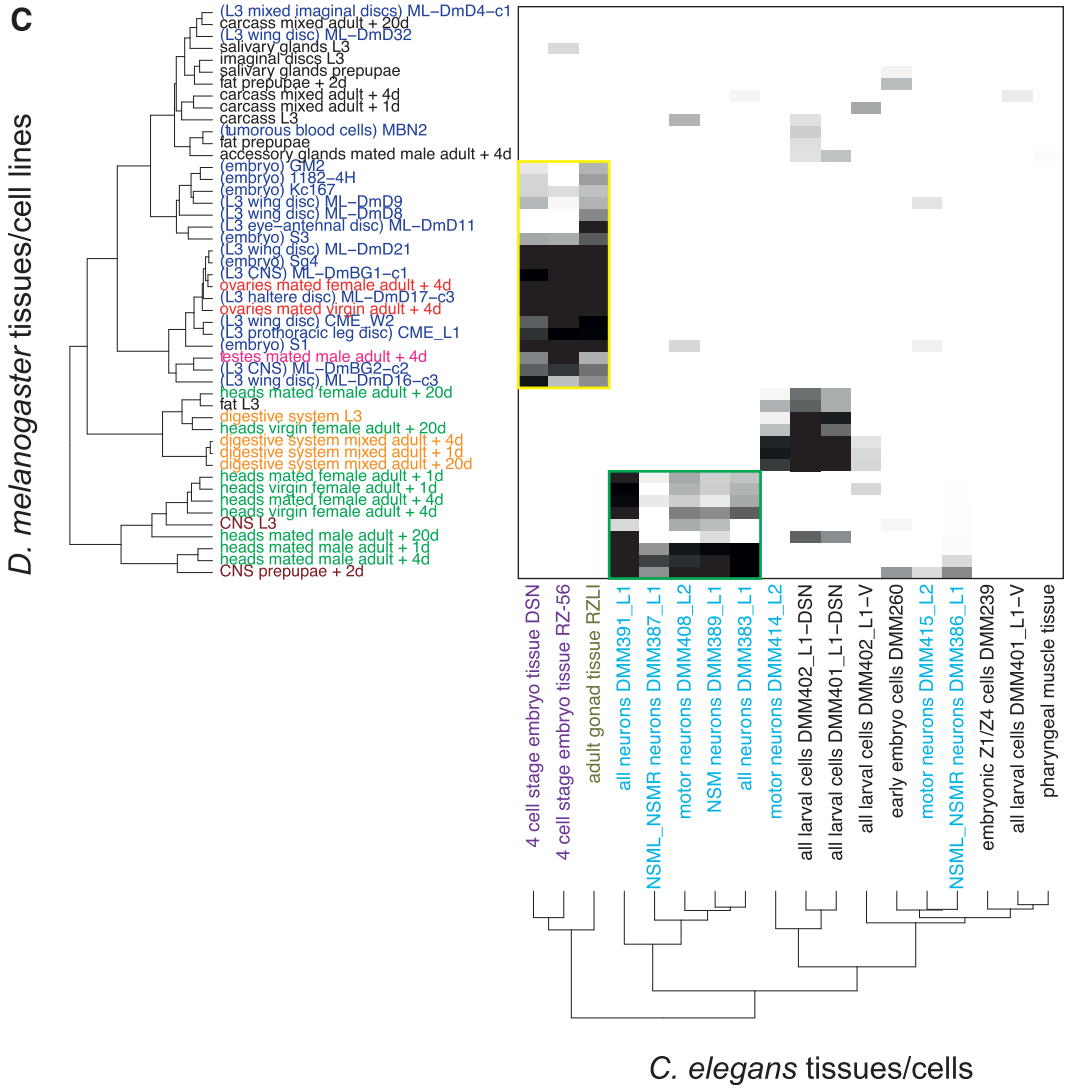


Figure 6. Comparison results of different developmental stages, tissues, cell lines, and dissected cells between *D. melanogaster* and *C. elegans*. Mapping scores shown in the heatmaps are $-\log_{10}$ transformed Bonferroni corrected *P*-values, which are calculated from the hypergeometric test in Figure 2B. (*A*) Comparison between *D. melanogaster* developmental stages and *C. elegans* tissues/dissected cells. (*B*) Comparison between *D. melanogaster* tissues/cell lines and *C. elegans* developmental stages. (*C*) Comparison between *D. melanogaster* tissues/cell lines and *C. elegans* tissues/dissected cells. Hierarchical clustering was applied to order the *D. melanogaster* tissues/cell lines and *C. elegans* tissues/dissected cells in *A*, *B*, and C. Tissues/dissected cells with similar origins and cell lines are marked with colors. For detailed information on the stage/tissue/cell labels, please refer to Supplemental Table S2.

noisier stage mapping patterns than our previous results in Figures 3A and 4A. An enhanced off-diagonal matching pattern of fly stages is seen in Supplemental Figure S13A compared to Figure 3A, suggesting the existence of fly genes with similar functions being expressed at those matched stages. This is consistent with a duplicated developmental program in fly, which is seen in the two-diagonal stage-mapping pattern between fly and worm (Fig. 5A). However, the between-species comparison results based on gene ontology (Supplemental Fig. S13C) only contain scarce mapping signals, probably due to the dissimilarity of orthologous genes' functions and the discrepancy in GO vocabulary usage for the two species (Aranguen et al. 2007).

Stage comparison by correlation analysis

In addition to our new comparison approach, we also attempted using the more standard correlation analysis to compare the developmental stages within fly and worm. Correlation coefficients (Pearson or Spearman) are calculated between every pair of fly or worm stages, and the results are summarized in Supplemental Figure S14. However, neither Pearson nor Spearman correlation is found effective in describing the relationship of developmental stages within either species (see Discussion.)

Discussion

We have developed a new hypothesis testing approach to compare developmental stages and tissues/cells within and between widely diverged species based on transcriptome-wide protein-coding gene expression data. Our approach centers on (1) using orthologous genes to link different species, and (2) identifying stage/tissue/cell-associated genes to capture transcriptome characteristics of different developmental stages and tissues/cells. Within- and between-species comparison results demonstrate the greater effectiveness of our approach compared with the more straightforward correlation analysis.

Although correlation analysis is a standard approach and has been widely used in gene expression studies within fly and worm as well as in other species (Arbeitman et al. 2002; Spencer et al. 2011; Necsulea et al. 2014), we found neither Pearson nor Spearman correlation was an effective measure in comparing developmental stages within fly or worm. It is known that Pearson correlation lacks robustness to outliers and its value depends heavily on the accuracy of gene expression estimates. Although Spearman correlation is more robust to outliers, it cannot produce a clear stage-mapping pattern either, because the high expression of housekeeping genes across all developmental stages would inflate the correlation values. Unlike the correlation analysis, our associated-gene-based comparison approach does not use all genes but focuses on small subsets of genes that capture transcriptome characteristics in different samples. Genes whose expression has little variance across different stages are excluded, resulting in more concentrated information for stage comparison. Moreover, since our approach is based on the selected associated gene sets rather than absolute gene expression levels, it is more robust to noise and biases in gene expression estimates.

Using our approach, we provide the first comprehensive transciptome-level comparison of multiple developmental stages, tissues, and cells between *D. melanogaster* and *C. elegans*, and our study has revealed several connections (i.e., mappings) between developmental stages and tissues/cells both within and between the two species that were unknown to us. Most importantly, in the stage comparison between fly and worm, we found two parallel collinear patterns between their life cycles. One pattern covers early embryonic through larval stages in both species, and a second parallel pattern is formed between fly prepupal through adult stages and worm late embryonic through adult stages. This result most likely is due to the recapitulation of the *D. melanogaster* life cycle, where rounds of proliferation of undifferentiated cells are followed by differentiation during embryogenesis and also during larval and pupal stages. The two parallel patterns are largely consistent with the within-fly stage mapping results (Fig. 3A; Supplemental Fig. S13A), as the nonadjacent fly stages mapped to each other within fly are also mapped to the same worm stage. This further emphasizes the repetition of developmental programs in the fly life cycle. In addition, the expression of many-to-one fly-worm orthologs contributes to the parallel patterns, suggesting there has been more gene duplication and subfunctionalization in fly development than in worm.

Further analyses based on gene ontology (GO) provide functional support for us to better understand the underlying mechanisms that lead to these observed comparison results. However, direct incorporation of GO into the stage/tissue/cell comparison remains a difficult task due to the difference in orthologous genes' functions and the discrepancy between the GO vocabulary annotated for the two species (Aranguren et al. 2007). Moreover, since splicing regulation in transcription and translation play significant roles in an organism's development and cell/tissue differentiation (Barberan-Soler and Zahler 2008; Salomonis et al. 2010), it is also important to consider additional layers of regulation, including splicing and translation, in refining the comparison results of different stages, tissues, and cells.

Our statistical approach is also directly applicable to comparing large-scale biological samples in terms of gene expression dynamics in other biological contexts. It can add a new dimension to many existing comparative genomics studies, for example, the conservation of cell differentiation and development processes across vertebrates (Domazet-Lošo and Tautz 2010; Irie and Kuratani 2011). In addition to the advantages of our statistical approach in comparing and distinguishing biological samples, the associated-gene sets identified by our approach could also provide further biological insights.

Methods

Estimating gene expression in developmental stages and tissues/cells

Cufflinks (version 1.3.0, supplied with reference annotation, i.e., using "-G" option; Trapnell et al. 2010) was used to estimate the expression of 13,781 *D. melanogaster* protein-coding genes in 30 developmental stages, 29 tissues, and 19 cell lines, and the expression of 20,389 *C. elegans* protein-coding genes in 35 developmental stages, 4 tissues, and 14 dissected cells from modENCODE RNA-seq data sets mapped to fly and worm reference genomes (see Supplemental Table S2 for data links). A few but not all stages/tissues/cells have biological replicates, for which we merged the replicates into one data set, so that every stage/tissue/cell ended up with one mapped RNA-seq data set. Gene annotations are from Ensembl assembly 66 (i.e., BDGP 5.64 for *D. melanogaster* and WS 220 for *C. elegans*). All the gene expression estimates returned by Cufflinks are in FPKM (fragments per kilobase of transcript per million mapped reads) units. Hence, every fly and worm gene has one FPKM value per developmental stage/tissue/cell.

Identification of stage-/tissue-/cell-associated genes

We identified associated genes in a biological sample as the genes highly expressed in that sample relative to other samples. Here we use the identification of stage-associated genes for fly stages as an example. For every fly gene, suppose its expression estimates (in FPKM units) in the 30 developmental stages are $e_1, \ldots, e_{30}$. We normalized them as $z_1, \ldots, z_{30}$, where $z_i = \frac{e_i - \bar{e}}{s}$, $i = 1, \ldots, 30$ are the normalized Z-scores, and

$$\bar{e} = \frac{1}{30}\sum_{i=1}^{30} e_i$$

and

$$s = \sqrt{\frac{1}{29}\sum_{i=1}^{30}(e_i - \bar{e})^2}$$

are the mean and standard deviation of gene expression across the 30 stages. Note that e_i represents the absolute expression estimate of the gene at stage i, and z_i represents the relative expression estimate of the gene at stage i as compared to other stages. For example, for each fly stage, we would like to select the fly genes whose relative expression is high and whose absolute expression is distinguishable from background noise. The selection threshold used in this study is $z_i > 1.5$ and $e_i \geq 1$. The genes satisfying this threshold are selected as the associated genes of stage i. Based on our experience with the Cufflinks gene expression estimates from sequencing data with a similar number of reads per sample, FPKM = 1 is a reasonable cutoff to distinguish real gene expression signal from background noise. We tried two other thresholds on the relative gene expression: $z_i > 1.2$ and $z_i > 1.8$, and found the comparison results very robust to the three thresholds, suggesting that $z_i > 1.5$ is a reasonable threshold. Using Z-scores as a criterion, we are able to capture the genes (e.g., transcription factors) that are expressed at a relatively low level but specific to a particular time point or tissue.

For worm developmental stages, we used the same method and threshold to select their stage-associated genes. For fly tissues/cell lines (and also worm tissues/cells), we treated them in aggregate like developmental stages in selecting their tissue/cell-associated genes. Hence, for every fly/worm stage/tissue/cell, the protein-coding genes that are relatively highly expressed in it but not always highly expressed in other stages/tissues/cells are selected as its associated genes.

Hypergeometric testing in within-species stage/tissue/cell comparison

Given two stages, or a stage and a tissue/cell, or two tissues/cells of the same species (i.e., *D. melanogaster* or *C. elegans*), we compared them by testing the dependence of their associated genes, e.g., gene sets A and B. We regarded all the protein-coding genes of the species as the population, and regarded the associated-gene sets A and B as two samples drawn from the population. The null hypothesis to be tested against is that A and B are two independent samples from the population; the alternative hypothesis is that A and B are dependent samples. This becomes a standard hypergeometric test, and the test statistic is the number of genes shared by A and B. Given the sizes of A and B, the larger the test statistics is, the higher the likelihood that the null hypothesis will be rejected. The P-value of the test statistic is calculated as

$$P = \sum_{i=|A\cap B|}^{\min(|A|,|B|)} \frac{\binom{n}{i}\binom{n-i}{|A|-i}\binom{n-|A|}{|B|-i}}{\binom{n}{|A|}\binom{n}{|B|}},$$

where n is the total number of protein-coding genes, and $|A|, |B|$ and $|A \cap B|$ are the numbers of genes in gene sets A, B, and $A \cap B$. For example, out of 13,781 fly protein-coding genes, fly stages "Embryo 0–2h" and "Adult Female + 5d" each have 3012 and 1102 associated genes, of which 864 genes are shared. That is, $n = 13{,}781$, $|A| = 3012$, $|B| = 1102$, and $|A \cap B| = 864$. Then, the P-value $< 10^{-300}$.

Hence, for any two stages, or a stage and a tissue/cell, or two tissues/cells, the P-value indicates the level of their dependence, in other words, the strength of their mapping in the comparison result. Due to the multiple testing issue, we corrected the P-value by Bonferroni correction

$$\text{Bonferroni corrected } P\text{-value} = P\text{-value} \times (\#\text{ of pairwise comparison})$$

In the comparison of the 30 developmental stages within fly, the number of pairwise comparison is $30 \times 30 = 900$. We then defined the mapping score as

$$\text{mapping score} = -\log_{10}(\text{Bonferroni corrected } P\text{-value})$$

and summarized the mapping scores of all pairwise comparison into a matrix. If rows or columns of the matrix correspond to developmental stages, they will be ordered by the temporal order of the stages; otherwise, if rows or columns correspond to tissues/cells, they will be grouped by hierarchical clustering. The ordered matrix will then be presented by a heatmap (e.g., Figs. 3A–D, 4A–C) to illustrate the mapping patterns.

Hypergeometic testing in between-species stage/tissue/cell comparison

Given two stages, or a stage and a tissue/cell, or two tissues/cells from different species (i.e., *D. melanogaster* and *C. elegans*), we compared them by testing the dependence of their associated genes, e.g., fly gene set F and worm gene set W, in terms of orthology. We restricted F (and W) to the associated genes that have worm (and fly) orthologs. We used the 11,403 modENCODE ortholog pairs between the two species (Table 1; Supplemental Table S1; Wu et al. 2014) as the population, represented by a two-column array of 11,403 rows:

$$\begin{array}{ccc} \text{fly gene} & & \text{worm gene} \\ f_1 & \leftrightarrow & w_1 \\ \ldots & & \ldots \\ f_{11,403} & \leftrightarrow & w_{11,403} \end{array},$$

where f_i and w_i are the fly and worm genes in the i-th ortholog pair. Please note that there exist repetitive genes in $\{f_1, \ldots, f_{11,403}\}$ and $\{w_1, \ldots, w_{11,403}\}$ due to the existence of one-to-many, many-to-one and many-to-many ortholog pairs. Since F and W contain no repetitive genes, we defined $F' = \{f_i\text{: if } f_i \in F, i = 1, \ldots, 11{,}403\} \subset \{f_1, \ldots, f_{11,403}\}$ and $W' = \{w_i\text{: if } w_i \in W, i = 1, \ldots, 11{,}403\} \subset \{w_1, \ldots, w_{11,403}\}$ as alternative versions of F and W with repetitive genes. In other words, F' corresponds to a subset of the left column in the array above; of the 11,403 ortholog pairs, if the fly genes are in F, the left-hand side will be in F'. W' is generated in a similar way. We then regarded F' as a sample from $\{f_1, \ldots, f_{11,403}\}$ (i.e., the fly gene part of the population) and W' as a sample from $\{w_1, \ldots, w_{11,403}\}$ (i.e., the worm gene part of the population). Because of the one-to-one projection relationship between $\{f_1, \ldots, f_{11,403}\}$ and $\{w_1, \ldots, w_{11,403}\}$, we can consider F' and W' as two samples from the same population.

The null hypothesis to be tested against is that F' and W' are independent samples from the population; the alternative hypothesis is that F' and W' are dependent samples. This becomes a hypergeometric test setting, and the test statistic is the number of

ortholog pairs existing between F' and W', defined as T. The larger the test statistics, the higher the likelihood that the null hypothesis will be rejected. The P-value of the test statistic is calculated as

$$P = \sum_{i=T}^{\min(|F'|,|W'|)} \frac{\binom{11,403}{i}\binom{11,403-i}{|F'|-i}\binom{11,403-|F'|}{|W'|-i}}{\binom{11,403}{|F'|}\binom{11,403}{|W'|}},$$

where $|F'|$ and $|W'|$ are the numbers of elements (including repetitive genes) in gene sets F' and W'. For example, fly stage "Embryo 0–2h" and worm stage "Young Adult" each have 1625 and 732 genes, of which 372 ortholog pairs exist. They correspond to $|F| = 1625$, $|F'| = 2274$, $|W| = 732$, $|W'| = 1385$, and $T = 372$. Then, the P-value $< 10^{-10}$.

Hence, for any two stages, or a stage and a tissue/cell, or two tissues/cells from two different species, the P-value indicates the extent of their dependence, in other words, the level of their mapping in the between-species comparison. Similar to the within-species comparison, we addressed the multiple testing issue by correcting the P-values with Bonferroni correction and subsequently calculated mapping scores as $-\log_{10}$(Bonferroni corrected P-value). The comparison result is also summarized in a matrix, in which hierarchical clustering is applied to order the tissues/cells as rows or columns, and finally represented by a heatmap (Figs. 5A, 6A–C). We note that we obtained a very similar comparison result by using the TreeFam (Li et al. 2006) orthologs (Supplemental Fig. S15), showing that our approach is robust to the choice of ortholog databases.

Gene functional analysis

We calculated the enrichment of a biological process (BP) gene ontology (GO) term in a biological sample (i.e., stage/tissue/cell) by a hypergeometric testing approach. Suppose there are N protein-coding genes in total (i.e., the population), out of which n genes are annotated with the GO term, and there are M associated genes in the biological sample, out of which m genes are annotated with the GO term. The null hypothesis is that the enrichment level of the GO term in the sample-associated genes is the same as that in the gene population. The alternative hypothesis is that the enrichment level of the GO term in the sample-associated genes is greater than that in the gene population. Under the null hypothesis, the number of sample-associated genes that are annotated with the GO term should follow a hypergeometric distribution, and a P-value can be calculated as

$$P = \sum_{x=m}^{M} \frac{\binom{n}{x}\binom{N-n}{M-x}}{\binom{N}{M}}.$$

A smaller P-value means that the GO term has greater enrichment in the biological sample. We plotted the enrichment patterns for a few selected top enriched GO terms across either all fly/worm developmental stages or all tissues/cells. For fly developmental stages and tissues/cells, we selected the GO terms whose P-values (uncorrected for multiple testing issue) are smaller than 10^{-7}, resulting in 236 GO terms in Supplemental Figure S5 and 229 terms in Supplemental Figure S6. A subset of these GO terms that are enriched in specific fly stages, tissues, and cells is shown in Supplemental Figure S4. For worm developmental stages and tissues/cells, we selected the GO terms whose uncorrected P-values are smaller than 10^{-6}, resulting in 118 GO terms in Supplemental Figure S9 and 69 terms in Supplemental Figure S10. A subset of these GO terms that are enriched in specific worm stages, tissues, and cells is shown in Supplemental Figure S8. Supplemental Figures S4–S6 and S8–S10 plot the enrichment patterns of the selected GO terms as heatmaps, with the enrichment score as $-\log_{10}$(Bonferroni corrected P-value), in which the Bonferroni corrected P-value = P-value × (# of selected GO terms) × (# of stages or tissues/cells).

Hypothesis testing approach for stage comparison based on gene ontology

We used similar hypothesis testing approaches to compare developmental stages within fly/worm and between the two species based on gene ontology instead of gene orthology. Given two stages, either from the same species or from different species, we first "translated" their associated genes into their corresponding leaf node biological process (BP) gene ontology (GO) terms, e.g., GO term sets A and B. We regarded all the leaf node BP GO terms corresponding to any fly or worm genes as the population, and regarded the GO term sets A and B as two samples drawn from the population. Then we used the same hypergeometric testing approach as in the within-species stage comparison but instead based on gene ontology, and summarized $-\log_{10}$(Bonferroni corrected P-value) as mapping scores in Supplemental Figure S13.

Acknowledgments

This work was supported by NIH/NHGRI U01 HG004271 (to Dr. Susan Celniker) and NIH/NHGRI RC2 HG005639 (to Dr. Manolis Kellis). Both grants are related to the modENCODE Project. This work was also supported by NIH/NHGRI U01 HG007031 to Dr. Peter J. Bickel. We would like to thank the modENCODE Consortium for their data and support, specifically Dr. Roger Hoskins, Dr. Robert Waterston, Dr. LaDeana Hillier, Dr. Sue Celniker, Dr. James B. Brown, Dr. Mark Gerstein, Dr. Angela Brooks, and Courtney French for their insightful comments. We would also like to thank Dat Duong for his work on the stage mapping by gene ontology. We are especially thankful for Dr. Mark Biggin's generous help on the gene functional analysis. Finally, we thank the three anonymous reviewers for their extensive comments that have greatly helped us improve this manuscript.

References

Adoutte A, Balavoine G, Lartillot N, Lespinet O, Prud'homme B, de Rosa R. 2000. The new animal phylogeny: reliability and implications. *Proc Natl Acad Sci* **97:** 4453–4456.

Aranguren ME, Bechhofer S, Lord P, Sattler U, Stevens R. 2007. Understanding and using the meaning of statements in a bio-ontology: recasting the Gene Ontology in OWL. *BMC Bioinformatics* **8:** 57.

Arbeitman MN, Furlong EE, Imam F, Johnson E, Null BH, Baker BS, Krasnow MA, Scott MP, Davis RW, White KP. 2002. Gene expression during the life cycle of *Drosophila melanogaster*. *Science* **297:** 2270–2275.

Ashburner M, Golic KG, Hawley RS. 2005. *Drosophila: a laboratory handbook*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Barberan-Soler S, Zahler AM. 2008. Alternative splicing regulation during *C. elegans* development: splicing factors as regulated targets. *PLoS Genet* **4:** e1000001.

Betschinger J, Knoblich JA. 2004. Dare to be different: asymmetric cell division in *Drosophila*, *C. elegans* and vertebrates. *Curr Biol* **14:** R674–R685.

Celniker SE, Dillon LAL, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Kellis M, Lai EC, Lieb JD, MacAlpine DM, et al. 2009. Unlocking the secrets of the genome. *Nature* **459:** 927–930.

Cherbas L, Willingham A, Zhang D, Yang L, Zou Y, Eads BD, Carlson JW, Landolin JM, Kapranov P, Dumais J, et al. 2011. The transcriptional diversity of 25 *Drosophila* cell lines. *Genome Res* **21:** 301–314.

Domazet-Lošo T, Tautz D. 2010. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* **468:** 815–818.

Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, et al. 2012. Ensembl 2012. *Nucleic Acids Res* **40:** D84–D90.

Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, Yip KY, Robilotto R, Rechtsteiner A, Ikegami K, et al. 2010. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science* **330:** 1775–1787.

Gerstein MB, Rozowsky J, Yan K-K, Wang D, Cheng C, Brown JB, Davis CA, Hillier L, Sisu C, Li JJ, et al. 2014. Comparative analysis of the transcriptome across distant species. *Nature* doi: 10.1038/nature13424.

Holter NS, Mitra M, Maritan A, Cieplak M, Banavar JR, Fedoroff NV. 2000. Fundamental patterns underlying gene expression profiles: simplicity from complexity. *Proc Natl Acad Sci* **97:** 8409–8414.

Irie N, Kuratani S. 2011. Comparative transcriptome analysis reveals vertebrate phylotypic period during organogenesis. *Nat Commun* **2:** 248.

Jiang M, Ryu J, Kiraly M, Duke K, Reinke V, Kim SK. 2001. Genome-wide analysis of developmental and sex-regulated gene expression profiles in *Caenorhabditis elegans*. *Proc Natl Acad Sci* **98:** 218–223.

Kalinka AT, Tomancak P. 2012. The evolution of early animal embryos: conservation or divergence? *Trends Ecol Evol* **27:** 385–393.

Kalinka AT, Varga KM, Gerrard DT, Preibisch S, Corcoran DL, Jarrells J, Ohler U, Bergman CM, Tomancak P. 2010. Gene expression divergence recapitulates the developmental hourglass model. *Nature* **468:** 811–814.

Kim SK, Lund J, Kiraly M, Duke K, Jiang M, Stuart JM, Eizinger A, Wylie BN, Davidson GS. 2001. A gene expression map for *Caenorhabditis elegans*. *Science* **293:** 2087–2092.

Lesch BJ, Page DC. 2012. Genetics of germ cell development. *Nat Rev Genet* **13:** 781–794.

Lettre G, Hengartner MO. 2006. Developmental apoptosis in *C. elegans*: a complex CEDnario. *Nat Rev Mol Cell Biol* **7:** 97–108.

Li H, Coghlan A, Ruan J, Coin LJ, Hériché J-K, Osmotherly L, Li R, Liu T, Zhang Z, Bolund L, et al., 2006. TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res* **34:** D572–D580.

Lott SE, Villalta JE, Schroth GP, Luo S, Tonkin LA, Eisen MB. 2011. Noncanonical compensation of zygotic X transcription in early *Drosophila melanogaster* development revealed through single-embryo RNA-seq. *PLoS Biol* **9:** e1000590.

Marygold SJ, Roote J, Reuter G, Lambertsson A, Ashburner M, Millburn GH, Harrison PM, Yu Z, Kenmochi N, Kaufman TC, et al. 2007. The ribosomal protein genes and *Minute* loci of *Drosophila melanogaster*. *Genome Biol* **8:** R216.

The modENCODE Consortium, Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, et al. 2010. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* **330:** 1787–1797.

Montell DJ. 1999. The genetics of cell migration in *Drosophila melanogaster* and *Caenorhabditis elegans* development. *Development* **126:** 3035–3046.

Nayak S, Goree J, Schedl T. 2005. *fog-2* and the evolution of self-fertile hermaphroditism in *Caenorhabditis*. *PLoS Biol* **3:** 57–71.

Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Grützner F, Kaessmann H. 2014. The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505:** 635–640.

Pearson JC, Lemons D, McGinnis W. 2005. Modulating Hox gene functions during animal body patterning. *Nat Rev Genet* **6:** 893–904.

Salomonis N, Schlieve CR, Pereira L, Wahlquist C, Colas A, Zambon AC, Vranizan K, Spindler MJ, Pico AR, Cline MS, et al. 2010. Alternative splicing regulates mouse embryonic stem cell pluripotency and differentiation. *Proc Natl Acad Sci* **107:** 10514–10519.

Spencer WC, Zeller G, Watson JD, Henz SR, Watkins KL, McWhirter RD, Petersen S, Sreedharan VT, Widmer C, Jo J, et al. 2011. A spatial and temporal map of *C. elegans* gene expression. *Genome Res* **21:** 325–341.

Stolc V, Gauhar Z, Mason C, Halasz G, van Batenburg MF, Rifkin SA, Hua S, Herreman T, Tongprasit W, Barbano PE, et al. 2004. A gene expression map for the euchromatic genome of *Drosophila melanogaster*. *Science* **306:** 655–660.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. 2010. Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* **28:** 511–515.

Weigmann K, Klapper R, Strasser T, Rickert C, Technau G, Jäckle H, Janning W, Klämbt C. 2003. FlyMove–a new way to look at development of *Drosophila*. *Trends Genet* **19:** 310–311.

Whitfield ML, Sherlock G, Saldanha AJ, Murray JI, Ball CA, Alexander KE, Matese JC, Perou CM, Hurt MM, Brown PO et al. 2002. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol Biol Cell* **13:** 1977–2000.

Wolpert L. 2011. *Principles of development*. Oxford University Press, Oxford, UK.

Wu YC, Bansal MS, Rasmussen MD, Herrero J, Kellis M. 2014. Phylogenetic identification and functional characterization of orthologs and paralogs across human, mouse, fly, and worm. *bioRxiv* doi: http://dx.doi.org/10.1101/005736.

Received November 24, 2013; accepted in revised form May 14, 2014.