Comparative analysis across ENCODE, modENCODE and mouseENCODE reveals hourglass patterns of developmental gene co-expression networks

Daifeng Wang[1,2,†], Fei He[3,†], Gang Fang[4], Koon-Kiu Yan[6,7], Jinrui Xu[6,7], Joel Rozowsky[6,7], Shuang Liu[6,7], Sergei Maslov[3], Mark Gerstein[6,7,8*]

[1]Department of Biomedical Informatics, Stony Brook University, Stony Brook, NY, USA; [2]Stony Brook Cancer Center, Stony Brook Medicine, Stony Brook, NY, USA; [3]Department of Plant Pathology, Kansas State University, Manhattan, KS, USA; [4]New York University Shanghai, China; [5]Department of Bioengineering, Carl R. Woese Institute for Genomic Biology, and National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Champaign, IL, USA; [6]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA; [7]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT, USA; [8]Department of Computer Science, Yale University, New Haven, CT, USA

*To whom correspondence should be addressed: pi@gersteinlab.org

[†]Those authors contributed equally

Abstract

The large genomic projects such as ENCODE, modENCODE and mouseENCODE provide large-scale datasets and enable the comparative genomic analysis across multiple organisms to discover the evolutionary genomic functions, especially for development. However, the comparative data integration and analysis across ENCODE related projects is still a challenge. To this end, this paper demonstrated a case study using integrated mod/mouse/ENCODE datasets for discovering the hourglass patterns of developmental gene co-expression network structures. The canonical hourglass behaviors have previously been observed at gross morphological and individual gene transcriptomic levels during embryogenesis, with the largest constraint occurring at the phylotypic stage (the "pinch" of the hourglass). Beyond this, we further clustered integrated RNA-seq data into the gene co-expression modules during embryogenesis for worm, fly and mouse, and found that their temporal interconnectivity during development has a 'network' hourglass pattern; i.e., cross-species conserved modules, rather than species-specific ones, achieve their highest network modularity (e.g., module's preservation degree) near before and at the particular developmental; e.g., the phylotypic stage of worm and fly, suggesting that various conserved functions start to become activated during the middle rather than the early or late embryonic stages. Furthermore, the ChIP-seq data in mod/ENCODE revealed that the transcription factors potentially regulating some of those modules are up-regulated at the onset of phylotypic stage. Finally, we also observed that the mouse and zebrafish orthologous genes of the conserved modules have the hourglass behaviors during their embryogenesis.

1. Introduction

Comparative genomics enables identifying the evolutionarily conserved and species-specific genomic elements across multiple organisms. However, systematic identification of the genomic elements on the genome wide, especially to advance the knowledge on functional genomics in mammals such as human and mouse is still challenge. To address this, a few large scientific consortia including ENCODE, mouseENCODE and modENCODE have systematically generated the large-scale next generation sequencing data (e.g., RNA-seq and ChIP-seq) for detecting the genomic and transcriptomic activities across human and model organisms (mouse, worm, fly) [1, 2, 3]. These cross-species datasets enable large-scale comparative genomic analysis to

Style Definition ... [1]
Deleted: transcriptomic network …nalysis across ... [2]
Deleted: Maslov[3,5]
Deleted: [3]Biology Department, Brookhaven National ... [3]
Deleted: Hourglass
Formatted ... [4]
Formatted ... [5]
Deleted: single-
Formatted ... [6]
Deleted: transcriptome
Formatted ... [7]
Deleted: In
Formatted ... [8]
Deleted: paper
Deleted: also found developmental hourglass patterns ... [10]
Formatted ... [9]
Formatted ... [11]
Deleted: orthologous genes between worm (*C.*
Moved down [1]: elegans) and fly (D. mel)
Formatted ... [12]
Deleted: based on the correlations of
Formatted ... [13]
Deleted: gene expression profiles
Formatted ... [14]
Deleted: embryonic
Formatted ... [15]
Deleted: . Some modules exist in both two organisms (
Formatted ... [16]
Deleted: .
Formatted ... [17]
Deleted: module), and others are more
Formatted ... [18]
Deleted: . We found that the conserved modules
Formatted ... [19]
Deleted: i.
Formatted ... [20]
Deleted: Coincidentally,
Formatted ... [21]
Deleted: We also found that the conserved modules a ... [22]
Formatted ... [23]
Deleted: Zebrafish
Formatted ... [24]
Deleted: at the gene network level
Formatted ... [25]
Deleted: the Zebrafish
Formatted ... [26]
Deleted: Nearly 200 years ago, Haeckel proposed the ... [27]

identify the conserved and species-specific genomic functions, especially for embryonic development. Thus, in this paper, we integrated these ENCODE-related datasets, performed the comparative analysis for the developmental gene co-expression networks across species, and found the network hourglass patterns that associate with cross-species conserved functions in embryogenesis.

Nearly 200 years ago, Haeckel proposed the recapitulation theory that the embryogenesis of animals resembles the successive evolutionary path from their ancestors[4]. The limited microscopic resolutions at that time did not enable biologists to gain a clear view of early embryogenesis. Before gastrulation, embryos from different animals look more different than they appear in later stages. The so-called 'ontogeny recapitulates phylogeny' is not accepted by modern biology[5]. However, the idea behind this theory persisted and shaped our understanding of development[6]. Currently, it is generally accepted that animals of the same phylum share a common morphological stage; i.e. the phylotypic stage during embryogenesis[7]. An 'hourglass' model was proposed to explain the existence of this conserved stage[8, 9]. Raff argued that the molecular signaling between different developmental modules (e.g., limbs) is extensive and highly inter-dependent at this stage. Any mutation in the genes that are functional during this time period may lead to fatality, thereby rendering it conserved across different animals[9]. In order to find experimental evidence to support this hypothetic mechanism, homologous traits between different animals were quantitatively measured and compared[10, 11, 12, 13]. This type of study was difficult because there was no universal standard to define homologous traits. Therefore, the proposed mechanism behind the hourglass behavior remains inconclusive[6].

The availability of genome-wide gene expression data allows us to study developmental processes at the molecular level. The divergence of gene expression follows an hourglass-like pattern in six Drosophila species, which have diverged over a course of 40 million years. The time-series microarray data of each species were first collected, and the smallest divergence of gene expression appeared at the mid-embryonic stage[14]. In addition to directly comparing gene expression, measuring the evolutionary age of a transcriptome also demonstrated that the mid-embryonic stage expresses more ancient genes than earlier or later stages[15, 16]. The hourglass-like pattern of conservation (in terms of conserved gene expression levels) holds true between different animals[17] and even between different phyla[3]. Those studies generally reveal that an hourglass pattern exists with respect to conserved gene expression[18].

Raff argued that the inter-dependent molecular signaling between different developmental modules is the main reason for a conserved middle stage[19]. Numerous studies tested this hypothetic mechanism using molecular experimental data. However, those tests were not focused on the modules or the interaction between them[6]. The module in Raff's proposal can be considered as organs, such as limb, which consists of a group of discrete cells[19]. This modularity also exists among the gene regulatory networks[20]. A recent study analyzed the gene co-expression modules during each stage of zebrafish embryogenesis and found the expression of genes within each module is most similar to their mouse orthologous genes at the early stages of embryogenesis[21], which however did not study the interactions between various modules during embryonic development. In this paper, in order to test Raff's hypothetic mechanism of the hourglass model, we used gene co-expression modules during embryogenesis that had been detected in our previous study to represent the developmental module. In particular as shown in Figure 1, we analyzed the conservations of gene co-expression connectivity for these modules

Deleted: (Irie and Kuratani, 2014).

Deleted: (Raff, 2007). This modularity also exists among the gene regulatory networks (Davidson and Erwin, 2006).

Deleted: (Piasecka, et al., 2013), which however did not study the interactions between various modules during embryonic development.

across developmental stages, and found that they also achieved the maximum conservations at the phylotypic stage. This represents a developmental hourglass pattern of developmental gene co-expression network structures, whereas our previous analysis revealed the hourglass patterns of modular expression differences; i.e., minimum expression level differences at the phylotypic stage.

2. Results

Gene regulation determines the attributes of an organism's phenotype, such as morphology, so conserved gene regulatory mechanisms controlling the developmental hourglass behaviors might exist. In this paper, we are interested in finding the gene regulatory patterns that drive developmental hourglass behaviors. It is known that if genes are co-expressed in a biological process, it is highly likely that they are all controlled by similar gene regulatory mechanisms[22]. Moreover, clustering the gene co-expression network into gene co-expression modules reveals the functional grouping of genes[23]. Thus, we use the gene co-expression network connectivity between and among various gene modules to represent the gene regulatory patterns. In addition, since we found that the orthologous genes have developmental hourglass behaviors, as well as conserved genomic functions, we first try to identify a set of evolutionarily conserved and species-specific gene modules from worm and fly developmental gene co-expression networks[3, 24], and then analyze their network characteristics to see if any hourglass patterns exist. In addition, we related these modules with mouse and Zebrafish data[2, 16], and also found the hourglass patterns during their embryogenesis.

2.1 Identification of conserved and species-specific gene modules between worm and fly during embryonic development

We used our recent cross-species clustering algorithm[21] to cluster worm and fly gene co-expression networks in embryonic development, and obtained 29 conserved gene modules that mainly consist of both worm and fly orthologous genes, 108 worm-specific gene modules and 52 fly-specific gene modules (see methods). The conserved gene modules have worm-fly orthologous genes with conserved functions. The species-specific gene modules contain the genes that have the functions specific to worm or fly (see Table S1).

We found that the enriched gene ontology terms of those gene modules indeed represent the conserved or species-specific functions. Here, we use worm gene modules as case studies. As shown in Figure 2, a conserved gene module (i.e. c4) is highly expressed around 3.5 hours after fertilization, when the zygotic genome forms[25]. It is not surprising that most of the genes within c4 are ribosomal genes (p-value = 0, Table S1), since huge volumes of proteins are synthesized during cell division. Another conserved gene module (c6) is only highly expressed at the beginning and then quickly down-regulated, which is a typical pattern of maternal gene expression (Figure 2)[26]. The 'proteasome complex' is over-represented in this gene module (p-value < 1e-9), which is consistent with the knowledge that maternal proteins need to be cleared during embryogenesis[27]. One should note that the gene modules mentioned here are conserved between distantly related species[3]. Unlike general gene co-expression modules in which genes are co-regulated, our modules contain genes that are also conserved between worm and fly. Those conserved gene modules very likely represent the basic components of embryogenesis[9, 20].

Two worm-specific gene modules were shown in Figure 2. The w10 is enriched with the gene ontology (GO) term 'sensory perception of chemical stimulus (p-value < 1e-10) and w101 is

Deleted: It is known that if genes are co-expressed in a biological process, it is highly likely that they are all controlled by similar gene regulatory mechanisms (Kim *et al.*, 2001). Moreover, clustering the gene co-expression network into gene co-expression modules reveals the functional grouping of genes (Stuart *et al.*, 2003). Thus, we use the gene co-expression network connectivity between and among various gene modules to represent the gene regulatory patterns. In addition, since we found that the orthologous genes have developmental hourglass behaviors, as well as conserved genomic functions, we first try to identify a set of evolutionarily conserved and species-specific gene modules from worm and fly developmental gene co-expression networks (Gerstein *et al.*, 2014), and then analyze their network characteristics to see if any hourglass patterns exist

Deleted: (Yan *et al.*, 2014)

Deleted: c4) is highly expressed around 3.5 hours after fertilization, when the zygotic genome forms (Tadros and Lipshitz, 2009). It is not surprising that most of the genes within c4 are ribosomal genes (p-value = 0, Table S1), since huge volumes of proteins are synthesized during cell division. Another conserved gene module (c6) is only highly expressed at the beginning and then quickly down-regulated, which is a typical pattern of maternal gene expression (Figure 2) (Baugh, 2003). The 'proteasome complex' is over-represented in this gene module (p-value < 10^{-10}), which is consistent with the knowledge that maternal proteins need to be cleared during embryogenesis (Du *et al.*, 2015). One should note that the gene modules mentioned here are conserved between distantly related species (Gerstein *et al.*, 2014). Unlike general gene co-expression modules in which genes are co-regulated, our modules contain genes that are also conserved between worm and fly. Those conserved gene modules very likely represent the basic components of embryogenesis (Davidson and Erwin, 2006; Raff, 2007).

Deleted: stimulus'

Deleted: 10

Formatted: Not Superscript/ Subscript

enriched with the GO term 'neuropeptide signaling pathway (p-value < 1e-7). Both show a gradually increased expression level during embryogenesis, indicating that the interaction between embryo and environment becomes more intensive as the embryo develops[28].

2.2 Conserved gene modules are highly inter-connected with each other at the mid-embryonic stage

As proposed by Raff in 1996, a developmental module should be able to interact with other developmental modules in a hierarchically organized and genetically discrete way[19]. A developmental module is an independent functional unit, such as a limb bud[19]. This definition of a module at the anatomical level can be leveraged to the partitioning of a developing embryo[29]. At the genetic level, a group of genes that are under the same regulatory control can also be considered to constitute a module[30], such as well-characterized protein complexes (e.g. ribosomes)[31]. Omics data are an ideal start for detecting those subcellular organizational patterns[32]. Using traditional mathematical methods, it is easy to detect groups of genes that are tightly connected with each other. Biological modules are usually enriched among those network clusters[33]. Raff argued the increased inter-connection between modules leads to the conservation of the phylotypic stage. Here, we use our gene modules to represent the organizational groups and want to check their inter-connections. Since these gene modules are measured by correlating their expression profiles during embryogenesis[3], the 'inter-connection' between modules can be measured by the co-expression degree; e.g., correlation between the eigengenes of two modules. The modular eigengene represents the temporal expression dynamic patterns of modular genes (Methods).

We calculated the correlation coefficients between modular eigengenes at different time periods of embryogenesis (Methods). For example, two conserved gene modules (c2 and c4) are most correlated around the time period containing 360 minutes after fertilization; i.e., the 12th time window, which coincides with the phylotypic stage[34] (Figure S1a). The c2 is enriched for the GO term 'transmembrane transporter activity' (p < 1e-16) while c4 is enriched for the term 'ribosome' (p < 2.2e-16). Although these two gene modules usually play a role independently, they seem to be under the same regulatory control during the worm phylotypic stage. This unusual increased correlation may lead to the hourglass pattern of development[19]. On the other hand, a pair of worm-specific gene modules (w10 and w13) show relatively low correlation during the phylotypic stage (Figure S1b), suggesting that species-specific gene modules may be under different regulatory controls at this stage. We further checked all pairwise correlations between conserved gene modules and worm-specific modules, respectively.

As shown in Figure 3a and Figure 4, the correlations between 29 conserved gene modules achieve their highest values at the phylotypic stage, which means Raff's proposed mechanism for the hourglass model can be observed using gene expression networks. However, the 108 worm-specific gene modules do not have an increased inter-connection during mid-embryogenesis (Figure 3b). Levin et al. showed that the distance between gene expression patterns between different worm species follows an hourglass-like pattern, where the most conserved expression patterns appeared during mid-embryogenesis[34]. Our analysis demonstrated that mid-embryogenesis also has the most inter-connections between different modules that are conserved between fly and worm. During the middle (phylotypic) stage, the conserved modules start to

Deleted: pathway'

Deleted: = 10

Formatted: Not Superscript/ Subscript

Deleted: (Perrimon *et al.*, 2012).

Deleted: A developmental module is an independent functional unit, such as a limb bud (Raff, 1996). This definition of a module at the anatomical level can be leveraged to the partitioning of a developing embryo (Reno *et al.*, 2008). At the genetic level, a group of genes that are under the same regulatory control can also be considered to constitute a module (Arnone and Davidson, 1997), such as well-characterized protein complexes (e.g. ribosomes) (Lacquaniti *et al.*, 2013). Omics data are an ideal start for detecting those subcellular organizational patterns (Barabási and Oltvai, 2004). Using traditional mathematical methods, it is easy to detect groups of genes that are tightly connected with each other. Biological modules are usually enriched among those network clusters (Zhu *et al.*, 2007). Raff argued the increased inter-connection between modules leads to the conservation of the phylotypic stage. Here, we use our gene modules to represent the organizational groups and want to check their inter-connections. Since these gene modules are measured by correlating their expression profiles during embryogenesis (Gerstein *et al.*, 2014), the 'inter-connection' between modules can be measured by the co-expression degree; e.g.,

Deleted: (Levin *et al.*, 2012) (Figure S1a). The c2 is enriched for the GO term 'transmembrane transporter activity' ($p = 10^{-16}$) while c4 is enriched for the term 'ribosome' ($p < 2.2x10^{-16}$).

Deleted: (Raff, 1996).

Deleted: (Levin *et al.*).

form due to the high modularity, but because they have to work together for conserved developmental functions, they retain high inter-connectivity.

2.3 Conserved gene modules showed highest preservation scores at the mid-embryonic stages from ahead to late of phylotypic stage

The classical definition of a biological module is usually an embryonic structure that has a clear morphological organization[35]. The early embryonic stage does not have this kind of individualization[36]. It is argued that early embryogenesis only contains a simple molecular network that lacks clear modularity[6]. While it is difficult to test this idea using empirical data, we can evaluate the modularity of our gene modules using WGCNA in different time periods of embryogenesis (see Methods). The modularity is calculated by a Z-score to represent the how well a gene module is preserved during a particular time period; i.e., a subset of time samples[37]. A Z-score higher than 4 generally represents a module is preserved, whereas Z-scores below 2 indicate that no module can be detected[37].

It is interesting to know whether the gene modules can be reproducibly detected at a specific stage of embryogenesis. Again, using a continuous time window of 6 time points (i.e., a time period of 3 hours), we calculated the preservation score (i.e. Z-score) for all gene modules for each time window. For example, the c1, a conserved gene module shows the highest expression abundance at the end of embryogenesis (Figure S2a), however, its preservation score is largest in the middle (Figure S2b), which covers the phylotypic stage and the stage ahead of it. The module c1 is enriched with the GO terms on cell-cell signaling ($p < 1.2e-15$). Since its preservation score becomes the highest near before and at the phylotypic stage, the associated biological functions are most activated during this period. On the other hand, a worm-specific gene module (w10), which is enriched with the GO term 'sensory perception of chemical stimulus' ($p < 1e-15$) shows relatively low preservation score during the phylotypic stage, although its expression abundance is relatively high during this period (Figure S3). Based on the observation of those two gene modules, we speculate that the activation of evolutionarily conserved gene modules may be associated with the phylotypic stage[19].

We further checked the preservation of all gene modules containing at least 50 genes during different time periods of embryogenesis. As expected, the conserved gene modules show the highest preservation score at mid-embryogenesis, which follows an hourglass-like pattern (Figure 5a). The worm-specific gene modules do not have this characteristic (Figure 5b), indicating that the hourglass pattern of embryo development is driven by evolutionarily conserved modules only. The high modularity at mid-embryogenesis suggests that the conserved modules start to form a module ahead of the phylotypic stage, and continue to work their functions during the phylotypic stage. After the conserved modules form early on of the phylotypic stage, their functions also should coordinate with each other at the phylotypic stage, which explains why the conserved modules have high inter-correlated eigengenes during the time periods covering the phylotypic stage (Figures 3 and 4).

In addition, we identified a group of TFs co-regulating the conserved modules and potentially drive the hourglass patterns. Because the genes in a same co-expression module are very likely co-regulated by similar gene regulatory programs, the high degree of preservation of multiple conserved gene co-expression modules at the middle embryonic stages imply that they are co-regulated specifically at mid-embryogenesis. As such, we identified potential transcription

Deleted: The classical definition of a biological module is usually an embryonic structure that has a clear morphological organization (Bolker, 2000). The early embryonic stage does not have this kind of individualization (Sulston *et al.*, 1983). It is argued that early embryogenesis only contains a simple molecular network that lacks clear modularity (Irie and Kuratani, 2014). While it is difficult to test this idea using empirical data, we can evaluate the modularity of our gene modules using WGCNA in different time periods of embryogenesis (see Methods). The modularity is calculated by a Z-score to represent the how well a gene module is preserved during a particular time period; i.e., a subset of time samples (Langfelder *et al.*, 2011). A Z-score higher than 4 generally represents a module is preserved, whereas Z-scores below 2 indicate that no module can be detected (Langfelder *et al.*, 2011). ... [28]

Deleted: = 1x10

Formatted: Not Superscript/ Subscript

Deleted: (Raff, 1996).

factors (TFs) regulating conserved modules from ChIP-seq data; i.e., they are found to have significantly a variety of target genes in conserved modules (See methods). For example, we found that five TFs (C04F5.9, CEH-90, DPL-1, F23B12.7 and MES-2), critical factors for embryonic development [38], co-regulate four conserved modules (c4, c7, c15 and c17). The DPL-1 is essential for the embryonic asymmetry (i.e. body plan). Three targeted gene modules of those TFs are enriched for 'embryo development' (p-value = 1.39e-40 for C4, 1.27e-3 for C7, 9.26e-5 for C15). As shown in Figure 6, these TFs are particularly upregulated at the beginning of the phylotypic stage (Fig. 6a), suggesting that they play potential regulatory roles driving the co-expression across these conserved modules at the phylotypic stage (Fig. 6b).

2.4 Conserved gene modules showed a specific hourglass pattern during early embryogenesis in mouse using mouseENCODE data

From an hourglass perspective, we analyzed the transcriptome data from early mouse embryonic development in mouseENCODE project[2], covering the stages from oocyte, zygote, early 2 cell, 2 cell, 4 cell to 8 cell. These clearly defined and well separated stages provide a unique opportunity to study hourglass pattern of conserved genes in early embryonic development within a short window. We identified 1,496 mouse genes that are orthologous to the conserved genes in the 16 modules between worm, fly and human[3]. For each module, the mouse genes are represented by the eigengene (Fig. 7a). The expressions of the 16 eigengenes have less variations at the 2 cell stages in term of standard deviations, whilst the expressions significantly diverge at both the earlier and later stages (Fig. 7b). This is probably because the 2 cell stages are critical to the formations of more complex patterns in the following stages, thus the conserved genes involved in pattern formations tend to be highly expressed (Fig. 7a), and as a result they are similar to each other. This is the typical cause of transcriptome responsible for canonical hourglass patterns with waists at phylotypic stages[14, 16, 39]. Here, we observed a special hourglass pattern with waist at 2 cell stages occurs much earlier than the canonical ones. Moreover, Pletikos et al observed a late hourglass pattern in human brain with the waist at infancy[40]. Taken together, these non-canonical hourglass patterns suggest that hourglass pattern in transcriptome may be a general sign, indicating the beginning of increase in developmental complexity and gaining functions.

3. Conclusion

In this paper, using the modENCODE expression datasets for organism development[3], we clustered orthologous genes between worm (C. elegans) and fly (D. mel) into gene co-expression modules based on the correlations of their temporal gene expression profiles during embryonic development. Some modules exist in both two organisms (i.e. conserved module), and others are more species-specific. Using those gene modules as an approximation to developmental modules, we tested the proposed hypothetical mechanism for the hourglass model[6, 19]. Our results support the notion that the conservation of the phylogenetic stage can be observed at the level of molecular networks. In details, we found that the conserved modules achieve their highest network modularity (i.e., module's preservation degree) near before and at the phylotypic stage, suggesting that various conserved functions start to become activated during the middle rather than the early or late embryonic stages. Coincidentally, the transcription factors potentially regulating some of those modules are up-regulated at the onset of phylotypic stage. We also found that the conserved modules are tightly inter-connected with each other near the phylotypic

Deleted: (Howe, et al., 2016),

Deleted: 39*10

Deleted: 27*10

Deleted: 26*10

Deleted: 3. Conclusion
Our previous work identified

Deleted: worm

Formatted: Font:Bold

Formatted: Font:Bold

Formatted: Font:Bold

Deleted: . Some modules are

Moved (insertion) [1]

Formatted: Font:Not Italic

Formatted: Font:Not Italic

Deleted: and fly, while

Deleted: (Raff, 1996; Irie and Kuratani, 2014).

stage, suggesting that the conserved functions should coordinate with each other at this middle stage. Thus, our results reveal that the multi-gene conserved modules follow the hourglass patterns in terms of their co-expression network connectivity in embryonic development. In contrast, we did not see such hourglass patterns from species-specific gene co-expression modules.

Embryo development is a cell differentiation process. The conserved gene modules are not yet formed at early stages based on our calculation of preservation (Figure 5). In later stages, the cells become differentiated and tissues/organs are relatively separated (these different tissues/organs are called 'modules' by developmental biologists). The expression data we measured is taken from a combination of all the cells. For example, if a gene is highly expressed in muscle but lowly expressed in skin, our data (based on the whole embryo) cannot catch such signals.

In addition to the worm and fly data, we also studied the hourglass pattern during the Zebrafish embryogenesis at the gene network level (See Methods). We found that there are 173 worm genes from our modules have a one-to-one orthologous gene in Zebrafish, and three worm-fly conserved modules contain at least 10 zebrafish orthologous genes; i.e., c4, c7 and c13. Surprisingly, the module c4 shows significant modularity score in the middle of zebrafish embryo development (Figure S4), indicating the modules defined by our study not only are conserved among distantly related species, but also achieves the highest modularity in the middle stage of zebrafish embryogenesis. The modules, c13 and c7 did not show significant modularity ($Z<4$) in any stages, however, they also show a marginal hourglass pattern (Figure S4).

In this paper, we studied the developmental gene co-expression networks that connect potentially co-regulated genes. Next generation sequencing data on gene regulation, including ChIP-seq and CLIP-seq, however, have directly provided the regulatory binding relationships between the gene regulatory factors and their target genes[41]. In addition, the developmental gene regulatory circuits were systematically discovered in simple organisms[20]. In the future, one can thus construct the developmental gene regulatory networks and try to discover the regulatory circuits that potentially drive the developmental hourglass patterns.

Deleted: Next generation sequencing data on gene regulation, including ChIP-seq and CLIP-seq, however, have directly provided the regulatory binding relationships between the gene regulatory factors and their target genes (Boyle *et al.*, 2014). In addition, the developmental gene regulatory circuits were systematically discovered in simple organisms (Davidson and Erwin, 2006).

4. Methods

4.1 Worm, fly and mouse gene expression data in embryonic development

The time-series gene expression data from worm and fly in embryonic development were generated by the modENCODE consortium using RNA-Seq[3]. The expression values from worm and fly were measured across 24 and 12 embyornic developmental stages, respectively. The total 10,031 worm-fly orthologous pairs (including one-to-one, one-to-many, many-to-many relationships from 5,769 unique worm orthologous genes and from 5,507 unique fly orthologous genes) between worm and fly were downloaded from the modENCODE website as they were compiled by the consortium[3]. In total, there are 20,377 worm genes and 13,623 fly genes. For each species, expression values in different developmental stages or cell lines were log-transformed and standardized and Spearman correlation coefficients were calculated for each pair of genes. For the mouse developmental data, we used the RNA-seq data from early mouse embryonic development in mouseENCODE project[2]. The data were collected at early

Deleted: fly

Deleted: (Gerstein *et al.*, 2014). The expression values from worm and fly were measured across 24 and 12 embyornic developmental stages, respectively. The total 10,031 worm-fly orthologous pairs (including one-to-one, one-to-many, many-to-many relationships from 5,769 unique worm orthologous genes and from 5,507 unique fly orthologous genes) between worm and fly were downloaded from the modENCODE website as they were compiled by the consortium (Gerstein *et al.*, 2014). In total, there are 20,377 worm genes and 13,623 fly genes. For each species, expression values in different developmental stages or cell lines were log-transformed and standardized and Spearman correlation coefficients were calculated for each pair of genes

developmental stages: oocyte, zygote, early 2 cell, 2 cell, 4 cell and 8 cell for the 1,496 orthologous mouse genes to worm and fly.

4.2 Conserved and species-specific gene co-expression modules

We constructed gene co-expression networks for worm and fly separately (nodes are genes, and edges connect genes if their spearman correlation coefficients exceed 0.9), and then applied OrthoClust to simultaneously cluster two networks to obtain the conserved and species-specific gene co-expression modules[24]. In total, we obtained 29 conserved gene modules that consist of both worm and fly genes, 108 worm-specific gene modules and 52 fly-specific gene modules.

4.3 Eigengenes of modules

The eigengene of a gene module is represented by the first right singular vector of singular value decomposition (SVD) of gene expression data matrix (genes by times) in this gene module, and is calculated using the svd() function in R. The expression value (at time t) of the eigengene in the i^{th} module is denoted as $m_i(t)$.

4.4 Selection of sliding windows

Each sliding window has six adjacent time points in worm embryo development. The k^{th} sliding window starts at the k^{th} time point, and ends at the $(k+5)^{th}$ time point in worm embryo development.

4.5 Correlations of modules

The correlation between gene modules i and j for the k^{th} sliding window, consisting of time points t_{k1}, t_{k2}, …, t_{k6} is calculated as $\boldsymbol{C}_k(i,j)$= Spearman correlation of two vectors: ($m_i(t_{k1})$, $m_i(t_{k2})$,..., $m_i(t_{k6})$) and ($m_j(t_{k1})$, $m_j(t_{k2})$,..., $m_j(t_{k6})$).

4.6 Calculating preservation score of modules using WGCNA

The preservation score of gene module was calculated using the modulePreservation function of R package, WGCNA[37]. For genes in a group, the average density and average connectivity were first computed. Using 100 randomized groups, the background distribution of those parameters was generated (i.e., a randomized group contains the same number of genes, which are randomly selected from the worm genome). Based on the background distribution, a Z-score can be determined. As recommended by the original authors, a module with a Z-score exceeding 4 means it can be reproducibly detected among different datasets[37]. Therefore, we used this Z-score as preservation score in our paper.

4.7 Identification of transcription factors (TFs) regulating gene co-expression modules

The potential target genes of transcription factors (TFs) are found if TFs have high binding signals at target gene promoter regions from TFs ChIP-seq experiments. The TFs regulating a gene co-expression module are the ones that have significantly numbers of target genes in the module (hypergeometric test $p<0.05$).

4.8 Zebrafish gene expression data in embryonic development

Formatted: Font color: Auto

Deleted: (Yan *et al.*, 2014).

Deleted: The preservation score of gene module was calculated using the modulePreservation package within WGCNA (Langfelder *et al.*, 2011). For genes in a group, the average density and average connectivity were first computed. Using 100 randomized groups, the background distribution of those parameters was generated (i.e., a randomized group contains the same number of genes, which are randomly selected from the worm genome). Based on the background distribution, a Z-score can be determined. As recommended by the original authors, a module with a Z-score exceeding 4 means it can be reproducibly detected among different datasets (Langfelder *et al.*, 2011). Therefore, we used this Z-score as preservation score in our pape ... [29]

The gene expression data of Zebrafish embryonic development were retrieved[16], and 26 samples before hatching were used in our analysis. The normalized expression values were download from NCBI GEO GSE24616. All gene expression values were log-10 transformed as suggested[21]. Replicates were then averaged. The one-to-one orthologous genes between zebrafish and worm were retrieved[42]. A time window of zebrafish embryogenesis covers a set of continuous 15 time points. The Z-score from WGCNA was used to measure the modularity of those zebrafish orthologous genes from our worm-fly conserved modules for each time window.

Deleted: from (Domazet-Loso and Tautz, 2010),

Deleted: in (Piasecka, et al., 2013).

Deleted: from (Cunningham, et al., 2015).

Figures

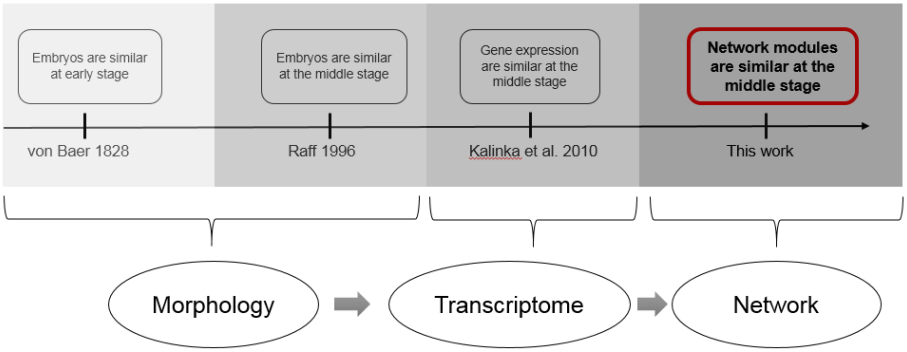

Figure 1. The history of developmental hourglass model. The concept that the early stage of different animals share similar characters was proposed in the early 19th centuries. In the 1990s, the developmental hourglass model was supported by modern technics. One hypothesis from Rudolf A. Raff attributed it to the complex molecular interactions in the middle stage of embryogenesis cells[9]. Recently, a series of work discovered that gene expression profiles of different animals are the most conserved at the phylotypic stage[14], supporting the hourglass model at the molecular level. In this work, we compared the gene co-expression modules for embryonic development between worm and fly, further supporting the hourglass model at the level of gene network.

Deleted: (Raff, 2007). Recently, a series of work, such as (Kalinka *et al.*, 2010),

Deleted: ,

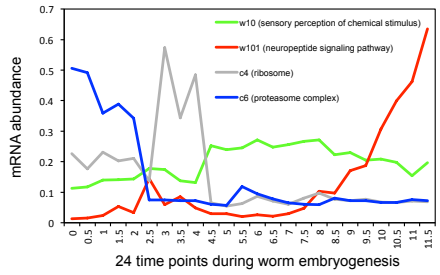

Figure 2. Expression profiles of selected gene modules. The w10 and w101 are two worm-specific gene modules, whereas c4 and c6 are two gene modules that are conserved between worm and fly. The representative enriched biological processes for each gene module are shown in the legend (see Supplemental Table 1 for detail). The eigengene of each gene module is used to represent the mRNA abundance (Y-axis). The X-axis represents the sampling time points (hours) of the RNAseq data.

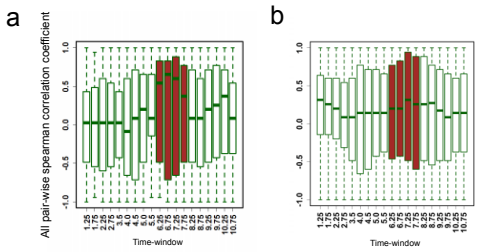

Figure 3. Correlation of expression profiles (eigengene) of gene modules during different time periods. All pairwise Spearman correlation coefficients among gene modules are shown in each time window of 3 hours during the worm embryogenesis for a) conserved gene modules and b) worm-specific gene modules. The red-colored boxes indicate the phylotypic stage. The Y-axis is the spearman correlation relationship.

Deleted: indicates

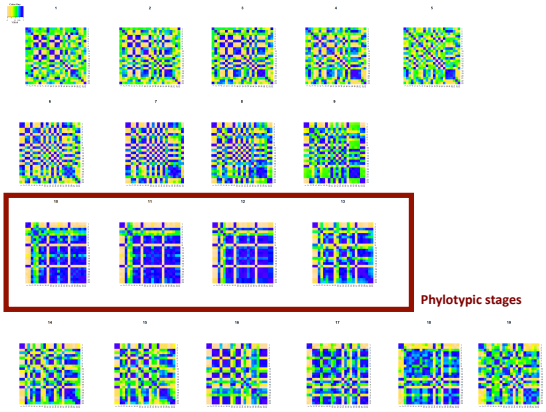

Figure 4. Similarity of expression profiles between different conserved gene modules in each time window of 3 hours during worm embryogenesis. As shown in the scale bar (top left), blue represents a positive correlation, yellow represents negative correlation, and green represents weak (i.e., close to 0) correlation. The time windows covering phylotypic stages are highlighted in brown boxes.

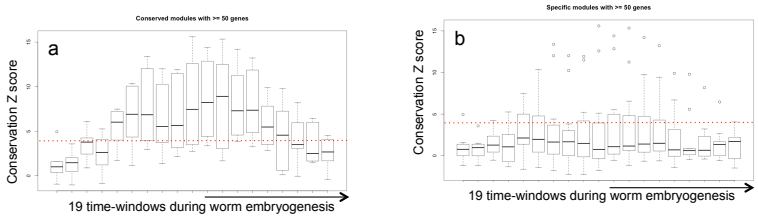

Figure 5. Preservation score among different time periods. Z-scores from 'modulePreservation' of WGCNA were used to evaluate preservation of gene modules. A Z-score exceeding 4 indicates the gene module can be detected. The X-axis represents time-windows (of 3 hours) during worm embryogenesis. a) conserved gene modules; b) worm-specific gene modules. Only modules with at least 50 genes are shown here.

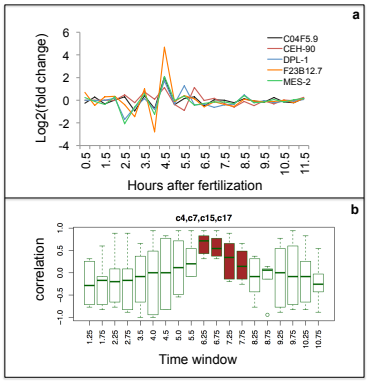

Figure 6. A case study of potential regulatory factors of conserved modules. Based on chip-seq data, the potential regulatory factors of each module were identified. Here, 4 conserved modules were significantly co-regulated by 5 TFs. (a) The expression pattern of TFs during embryogenesis, which was calculated as log2(fold change) between consecutive time points; (b) The correlation of expression profiles (i.e. eigengene) in each time window for 4 conserved modules.

Deleted:

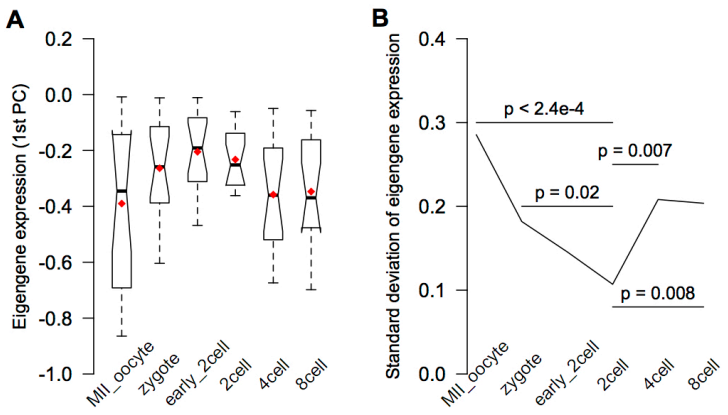

Figure 7. Mouse eigengene expressions and variations at early developmental stages. (A) the expression distributions of eigenegenes of the 16 modules. The red dots indicate average expressions. (B) the standard deviations of eigengene expressions of the 16 modules. The p-values are from F-statistics.

Supplemental materials

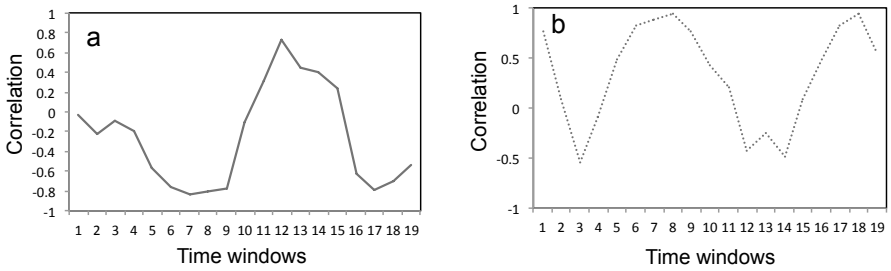

Figure S1 a) Correlation between a pair of conserved gene modules (c2 and c4) in different time periods; b) correlation between a pair of worm-specific gene modules (w10 and w13) in different time periods. The X-axis is the time window of 3 hours (including 6 sampling time points). The Y-axis is the Pearson correlation coefficient between the eigengene of a pair of gene modules.

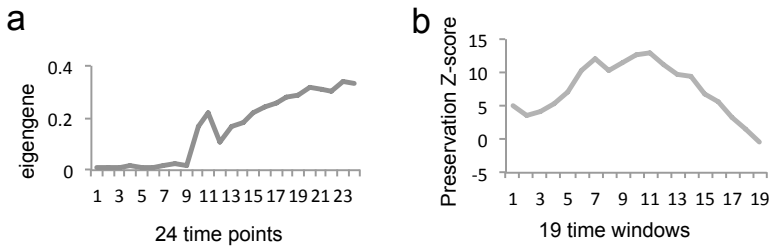

Figure S2. a) The expression profile of c1 during worm embryogenesis. The X-axis represents the 23 sampling time points. The Y-axis represents the eigengene of the gene module. b) The preservation score of c1 in different time windows of worm embryogenesis. The X-axis is the time windows of 6 sampling points. The Y-axis is the preservation score of the gene module in each time window.

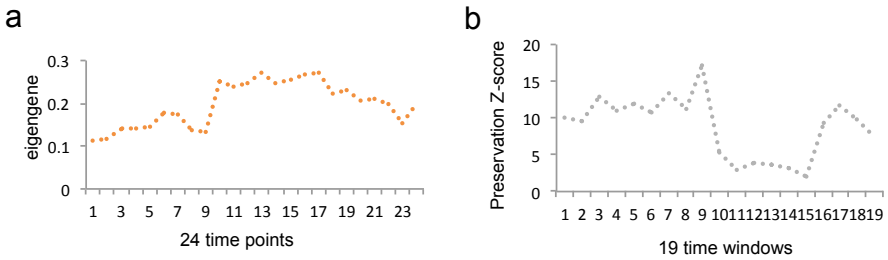

Figure S3. a) The expression profile of w10 during worm embryogenesis. The X-axis represents the 23 sampling time points. The Y-axis represents the eigengene of the gene module. b) The preservation of w10 in different time window of worm embryogenesis. The X-axis represents the time windows of 6 sampling points. The Y-axis represents the preservation score of the gene module in each time window.

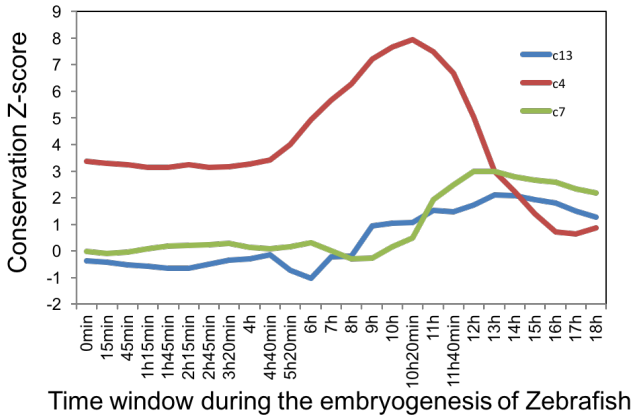

Figure S4. Preservation score among different time periods of the Zebrafish orthologs. The modularity of worm ortholog in Zebrafish embryogenesis was evaluated at different time window. Each time window contains 15 time points. The beginning of the time window was marked.

Table S1. The gene list and GO enrichment of each gene module.
We used Fisher's exact test followed by Benjamini–Hochberg correction to identify the enriched GO terms (FDR < 0.05). Only the most enriched terms are shown.

References

1. Consortium EP. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS biology* **9**, e1001046 (2011).

2. Yue F, *et al.* A comparative encyclopedia of DNA elements in the mouse genome. *Nature* **515**, 355-364 (2014).

3. Gerstein MB, *et al.* Comparative analysis of the transcriptome across distant species. *Nature* **512**, 445-448 (2014).

4. Hopwood N. *Haeckel's embryos : images, evolution, and fraud.*

5. Gould SJ. *Ontogeny and phylogeny*. Belknap Press of Harvard University Press (1977).

6. Irie N, Kuratani S. The developmental hourglass model: a predictor of the basic body plan? *Development* **141**, 4649-4655 (2014).

Deleted: ... [30]

7. British Society for Developmental Biology. Symposium (6th : 1982), Goodwin BC, Holder N, Wylie CC. *Development and evolution : the sixth symposium of the British Society for Developmental Biology*. Cambridge University Press (1983).

8. Duboule D. Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony. *Dev Suppl*, 135-142 (1994).

9. Raff RA. Written in stone: fossils, genes and evo-devo. *Nature reviews Genetics* **8**, 911-920 (2007).

10. Richardson MK, *et al.* There is no highly conserved embryonic stage in the vertebrates: implications for current theories of evolution and development. *Anat Embryol (Berl)* **196**, 91-106 (1997).

11. Galis F, Metz JA. Testing the vulnerability of the phylotypic stage: on modularity and evolutionary conservation. *J Exp Zool* **291**, 195-204 (2001).

12. Bininda-Emonds OR, Jeffery JE, Richardson MK. Inverting the hourglass: quantitative evidence against the phylotypic stage in vertebrate development. *Proc Biol Sci* **270**, 341-346 (2003).

13. Poe S, Wake MH. Quantitative tests of general models for the evolution of development. *Am Nat* **164**, 415-422 (2004).

14. Kalinka AT, *et al.* Gene expression divergence recapitulates the developmental hourglass model. *Nature* **468**, 811-814 (2010).

15. Quint M, Drost HG, Gabel A, Ullrich KK, Bonn M, Grosse I. A transcriptomic hourglass in plant embryogenesis. *Nature* **490**, 98-101 (2012).

16. Domazet-Loso T, Tautz D. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* **468**, 815-818 (2010).

17. Irie N, Kuratani S. Comparative transcriptome analysis reveals vertebrate phylotypic period during organogenesis. *Nature communications* **2**, 248 (2011).

18. Richardson MK. A phylotypic stage for all animals? *Developmental cell* **22**, 903-904 (2012).

19. Raff RA. *The shape of life : genes, development, and the evolution of animal form*. University of Chicago Press (1996).

20. Davidson EH. *The Regulatory Genome: Gene Regulatory Networks In Development And Evolution*, 1 edition edn. Academic Press (2006).

21. Piasecka B, Lichocki P, Moretti S, Bergmann S, Robinson-Rechavi M. The hourglass and the early conservation models--co-existing patterns of developmental constraints in vertebrates. *PLoS Genet* **9**, e1003476 (2013).

22. Kim SK, *et al.* A gene expression map for Caenorhabditis elegans. *Science* **293**, 2087-2092 (2001).

23. Stuart JM. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. *Science* **302**, 249-255 (2003).

24. Yan KK, Wang D, Rozowsky J, Zheng H, Cheng C, Gerstein M. OrthoClust: an orthology-based network framework for clustering data across multiple species. *Genome biology* **15**, R100 (2014).

25. Tadros W, Lipshitz HD. The maternal-to-zygotic transition: a play in two acts. *Development* **136**, 3033-3042 (2009).

26. Baugh LR, Hill AA, Slonim DK, Brown EL, Hunter CP. Composition and dynamics of the Caenorhabditis elegans early embryonic transcriptome. *Development* **130**, 889-900 (2003).

27. Du Z, He F, Yu Z, Bowerman B, Bao Z. E3 ubiquitin ligases promote progression of differentiation during C. elegans embryogenesis. *Developmental biology* **398**, 267-279 (2015).

28. Perrimon N, Pitsouli C, Shilo BZ. Signaling mechanisms controlling cell fate and embryonic patterning. *Cold Spring Harb Perspect Biol* **4**, a005975 (2012).

29. Reno PL, McCollum MA, Cohn MJ, Meindl RS, Hamrick M, Lovejoy CO. Patterns of correlation and covariation of anthropoid distal forelimb segments correspond to Hoxd expression territories. *J Exp Zool B Mol Dev Evol* **310**, 240-258 (2008).

30. Arnone MI, Davidson EH. The hardwiring of development: organization and function of genomic regulatory systems. *Development* **124**, 1851-1864 (1997).

31. Lacquaniti F, Ivanenko YP, d'Avella A, Zelik KE, Zago M. Evolutionary and developmental modules. *Front Comput Neurosci* **7**, 61 (2013).

32. Barabási A-L, Oltvai ZN. Network biology: understanding the cell's functional organization. *Nature Reviews Genetics* **5**, 101-113 (2004).

33. Zhu X, Gerstein M, Snyder M. Getting connected: analysis and principles of biological networks. *Genes Dev* **21**, 1010-1024 (2007).

34. Levin M, Hashimshony T, Wagner F, Yanai I. Developmental milestones punctuate gene expression in the Caenorhabditis embryo. *Developmental cell* **22**, 1101-1108 (2012).

35. Bolker JA. Modularity in Development and Why It Matters to Evo-Devo1. *American Zoologist* **40**, 770-776 (2000).

36. Sulston JE, Schierenberg E, White JG, Thomson JN. The embryonic cell lineage of the nematode Caenorhabditis elegans. *Developmental biology* **100**, 64-119 (1983).

37. Langfelder P, Luo R, Oldham MC, Horvath S. Is my network module preserved and reproducible? *PLoS computational biology* **7**, e1001057 (2011).

38. Howe KL*, et al.* WormBase 2016: expanding to enable helminth genomic research. *Nucleic Acids Res* **44**, D774-780 (2016).

39. Irie N, Sehara-Fujisawa A. The vertebrate phylotypic stage and an early bilaterian-related stage in mouse embryogenesis defined by genomic information. *BMC biology* **5**, 1 (2007).

40. Pletikos M*, et al.* Temporal specification and bilaterality of human neocortical topographic gene expression. *Neuron* **81**, 321-332 (2014).

41. Boyle AP, *et al.* Comparative analysis of regulatory information and circuits across distant species. *Nature* **512**, 453-456 (2014).

42. Cunningham F, *et al.* Ensembl 2015. *Nucleic Acids Res* **43**, D662-669 (2015).

Page 1: [1] Style Definition | Daifeng Wang | 12/22/17 10:20:00 PM

Default Paragraph Font

Page 1: [2] Deleted | Daifeng Wang | 12/22/17 10:20:00 PM

transcriptomic network

Page 1: [2] Deleted | Daifeng Wang | 12/22/17 10:20:00 PM

transcriptomic network

Page 1: [2] Deleted | Daifeng Wang | 12/22/17 10:20:00 PM

transcriptomic network

Page 1: [3] Deleted | Daifeng Wang | 12/22/17 10:20:00 PM

[3]Biology Department, Brookhaven National Laboratory, Upton, NY

Page 1: [4] Formatted | Daifeng Wang | 12/22/17 10:20:00 PM

Space After: 0 pt

Page 1: [5] Formatted | Daifeng Wang | 12/22/17 10:20:00 PM

Font color: Black

Page 1: [6] Formatted | Daifeng Wang | 12/22/17 10:20:00 PM

Font color: Black

Page 1: [7] Formatted | Daifeng Wang | 12/22/17 10:20:00 PM

Font color: Black

Page 1: [8] Formatted | Daifeng Wang | 12/22/17 10:20:00 PM

Font color: Black

Page 1: [9] Formatted | Daifeng Wang | 12/22/17 10:20:00 PM

Font color: Black

Page 1: [10] Deleted | Daifeng Wang | 12/22/17 10:20:00 PM

also found developmental hourglass patterns from the gene network structures. Using the modENCODE expression datasets for organism development, we

Page 1: [11] Formatted | Daifeng Wang | 12/22/17 10:20:00 PM

Font color: Black

Page 1: [12] Formatted | Daifeng Wang | 12/22/17 10:20:00 PM

Font:Not Italic

Page 1: [12] Formatted | Daifeng Wang | 12/22/17 10:20:00 PM

Font:Not Italic

Page 1: [12] Formatted	Daifeng Wang	12/22/17 10:20:00 PM

Font:Not Italic

Page 1: [12] Formatted	Daifeng Wang	12/22/17 10:20:00 PM

Font:Not Italic

Page 1: [13] Formatted	Daifeng Wang	12/22/17 10:20:00 PM

Font color: Black

Page 1: [14] Formatted	Daifeng Wang	12/22/17 10:20:00 PM

Font color: Black

Page 1: [15] Formatted	Daifeng Wang	12/22/17 10:20:00 PM

Font color: Black

Page 1: [16] Formatted	Daifeng Wang	12/22/17 10:20:00 PM

Font color: Black

Page 1: [17] Formatted	Daifeng Wang	12/22/17 10:20:00 PM

Font color: Black

Page 1: [18] Formatted	Daifeng Wang	12/22/17 10:20:00 PM

Font color: Black

Page 1: [19] Formatted	Daifeng Wang	12/22/17 10:20:00 PM

Font color: Black

Page 1: [20] Formatted	Daifeng Wang	12/22/17 10:20:00 PM

Font color: Black

Page 1: [20] Formatted	Daifeng Wang	12/22/17 10:20:00 PM

Font color: Black

Page 1: [20] Formatted	Daifeng Wang	12/22/17 10:20:00 PM

Font color: Black

Page 1: [20] Formatted	Daifeng Wang	12/22/17 10:20:00 PM

Font color: Black

Page 1: [21] Formatted	Daifeng Wang	12/22/17 10:20:00 PM

Font color: Black

Page 1: [22] Deleted	Daifeng Wang	12/22/17 10:20:00 PM

We also found that the conserved modules are tightly inter-connected with each other near the phylotypic stage, suggesting that the conserved functions should coordinate with each other at this middle stage. Thus, our results reveal that the multi-gene conserved modules follow the hourglass patterns in terms of their co-expression network connectivity in embryonic development. In contrast, we did not see such hourglass patterns from species-specific gene co-expression modules.

Page 1: [23] Formatted	Daifeng Wang	12/22/17 10:20:00 PM

Font color: Black

Page 1: [24] Formatted	Daifeng Wang	12/22/17 10:20:00 PM

Font color: Black

Page 1: [24] Formatted	Daifeng Wang	12/22/17 10:20:00 PM

Font color: Black

Page 1: [25] Formatted	Daifeng Wang	12/22/17 10:20:00 PM

Font color: Black

Page 1: [26] Formatted	Daifeng Wang	12/22/17 10:20:00 PM

Font color: Black

Page 1: [27] Deleted	Daifeng Wang	12/22/17 10:20:00 PM

Nearly 200 years ago, Haeckel proposed the recapitulation theory that the embryogenesis of animals resembles the successive evolutionary path from their ancestors (Hopwood). The limited microscopic resolutions at that time did not enable biologists to gain a clear view of early embryogenesis. Before gastrulation, embryos from different animals look more different than they appear in later stages. The so-called 'ontogeny recapitulates phylogeny' is not accepted by modern biology (Gould, 1977). However, the idea behind this theory persisted and shaped our understanding of development (Irie and Kuratani, 2014). Currently, it is generally accepted that animals of the same phylum share a common morphological stage; i.e. the phylotypic stage during embryogenesis (Sander, 1983). An 'hourglass' model was proposed to explain the existence of this conserved stage (Duboule, 1994; Raff, 1996). Raff argued that the molecular signaling between different developmental modules (e.g., limbs) is extensive and highly inter-dependent at this stage. Any mutation in the genes that are functional during this time period may lead to fatality, thereby rendering it conserved across different animals (Raff, 1996). In order to find experimental evidence to support this hypothetic mechanism, homologous traits between different animals were quantitatively measured and compared (Richardson *et al.*, 1997; Galis and Metz, 2001; Bininda-Emonds *et al.*, 2003; Steven Poe and Marvalee H. Wake, 2004). This type of study was difficult because there was no universal standard to define homologous

traits. Therefore, the proposed mechanism behind the hourglass behavior remains inconclusive (Irie and Kuratani, 2014).

The availability of genome-wide gene expression data allows us to study developmental processes at the molecular level. The divergence of gene expression follows an hourglass-like pattern in six Drosophila species, which have diverged over a course of 40 million years. The time-series microarray data of each species were first collected, and the smallest divergence of gene expression appeared at the mid-embryonic stage (Kalinka *et al.*, 2010). In addition to directly comparing gene expression, measuring the evolutionary age of a transcriptome also demonstrated that the mid-embryonic stage expresses more ancient genes than earlier or later stages (Domazet-Lošo and Tautz, 2010; Quint *et al.*, 2012). The hourglass-like pattern of conservation (in terms of conserved gene expression levels) holds true between different animals (Irie and Kuratani, 2011) and even between different phyla (Gerstein *et al.*, 2014). Those studies generally reveal that an hourglass pattern exists with respect to conserved gene expression (Richardson, 2012).

Raff argued that the inter-dependent molecular signaling between different developmental modules is the main reason for a conserved middle stage (Raff, 1996).

Page 5: [28] Deleted **Daifeng Wang** **12/22/17 10:20:00 PM**

The classical definition of a biological module is usually an embryonic structure that has a clear morphological organization (Bolker, 2000). The early embryonic stage does not have this kind of individualization (Sulston *et al.*, 1983). It is argued that early embryogenesis only contains a simple molecular network that lacks clear modularity (Irie and Kuratani, 2014). While it is difficult to test this idea using empirical data, we can evaluate the modularity of our gene modules using WGCNA in different time periods of embryogenesis (see Methods). The modularity is calculated by a Z-score to represent the how well a gene module is preserved during a particular time period; i.e., a subset of time samples (Langfelder *et al.*, 2011). A Z-score higher than 4 generally represents a module is preserved, whereas Z-scores below 2 indicate that no module can be detected (Langfelder *et al.*, 2011).

It is interesting to know whether the gene modules can be reproducibly detected at a specific stage of embryogenesis. Again, using a continuous time window of 6 time points (i.e., a time period of 3 hours), we calculated the preservation score (i.e. Z-score) for all gene modules for each time window. For example, the c1, a conserved gene module shows the highest expression abundance at the end of embryogenesis (Figure S2a), however, its preservation score is largest in the middle (Figure S2b), which covers the phylotypic stage and the stage ahead of it. The module c1 is enriched with the GO terms on cell-cell signaling ($p = 1.16x10^{-15}$).

Page 8: [29] Deleted **Daifeng Wang** **12/22/17 10:20:00 PM**

The preservation score of gene module was calculated using the modulePreservation package within WGCNA (Langfelder *et al.*, 2011). For genes in a group, the average density and average connectivity were first computed. Using 100 randomized groups, the background distribution of those parameters was generated (i.e., a randomized group contains the same number of genes, which are randomly selected from the worm genome). Based on the background distribution, a Z-score can be determined. As recommended by the original authors, a module with a Z-score

exceeding 4 means it can be reproducibly detected among different datasets (Langfelder *et al.*, 2011). Therefore, we used this Z-score as preservation score in our paper.

4.7 Identification of transcription factors (TFs) regulating gene co-expression modules

The potential target genes of transcription factors (TFs) are found if TFs have high binding signals at target gene promoter regions from TFs ChIP-seq experiments. The TFs regulating a gene co-expression module are the ones that have significantly numbers of target genes in the module (hypergeometric test $p<0.05$).

4.8 Zebrafish gene expression data in embryonic development

Page 14: [30] Deleted	Daifeng Wang	12/22/17 10:20:00 PM

Arnone,M.I. and Davidson,E.H. (1997) The hardwiring of development: organization and function of genomic regulatory systems. *Development*, **124**, 1851–1864.

Barabási,A.-L. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, **5**, 101–13.

Baugh,L.R. (2003) Composition and dynamics of the Caenorhabditis elegans early embryonic transcriptome. *Development*, **130**, 889–900.

Bininda-Emonds,O.R.P. *et al.* (2003) Inverting the hourglass: quantitative evidence against the phylotypic stage in vertebrate development. *Proc. Biol. Sci.*, **270**, 341–346.

Bolker,J.A. (2000) Modularity in Development and Why It Matters to Evo-Devo1. *Am. Zool.*, **40**, 770–776.

Boyle,A.P. *et al.* (2014) Comparative analysis of regulatory information and circuits across distant species. *Nature*, **512**, 453–456.

Davidson,E.H. and Erwin,D.H. (2006) Gene regulatory networks and the evolution of animal body plans. *Science*, **311**, 796–800.

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

Du,Z. *et al.* (2015) E3 ubiquitin ligases promote progression of differentiation during C. elegans embryogenesis. *Dev. Biol.*, **398**, 267–279.

Duboule,D. (1994) Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony. *Dev. Suppl.*, 135–142.

Galis,F. and Metz,J.A.J. (2001) Testing the vulnerability of the phylotypic stage: On modularity and evolutionary conservation. *J. Exp. Zool.*, **291**, 195–204.

Gerstein,M.B. *et al.* (2014) Comparative analysis of the transcriptome across distant species. *Nature*, **512**, 445–448.

Gould,S. (1977) Ontogeny and Phylogeny Harvard University Press.

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

Irie,N. and Kuratani,S. (2014) The developmental hourglass model: a predictor of the basic body plan. *Development*, **141**, 4649–4655.

Kalinka,A.T. *et al.* (2010) Gene expression divergence recapitulates the developmental hourglass model. *Nature*, **468**, 811–814.

Kim,S.K. *et al.* (2001) A gene expression map for Caenorhabditis elegans. *Science*, **293**, 2087–2092.

Lacquaniti,F. *et al.* (2013) Evolutionary and developmental modules. *Front. Comput. Neurosci.*, **7**, 61.
Langfelder,P. *et al.* (2011) Is my network module preserved and reproducible? *PLoS Comput. Biol.*, **7**, e1001057.
Levin,M. *et al.* (2012) Developmental Milestones Punctuate Gene Expression in the Caenorhabditis Embryo. *Dev. Cell*, **22**, 1101–1108.
Levin,M. *et al.* Supplemental Information Developmental Milestones Punctuate Gene Expression in the Caenorhabditis Embryo. **22**.
Perrimon,N. *et al.* (2012) Signaling mechanisms controlling cell fate and embryonic patterning. *Cold Spring Harb. Perspect. Biol.*, **4**, a005975.
Quint,M. *et al.* (2012) A transcriptomic hourglass in plant embryogenesis. *Nature*, **490**, 98–101.
Raff,R. (1996) The shape of life University of Chicago Press.
Raff,R.A. (2007) Written in stone: fossils, genes and evo-devo. *Nat Rev Genet*, **8**, 911–920.
Reno,P.L. *et al.* (2008) Patterns of correlation and covariation of anthropoid distal forelimb segments correspond to hoxd expression territories. *J. Exp. Zool. Part B Mol. Dev. Evol.*, **310**, 240–258.
Richardson,M.K. (2012) A Phylotypic Stage for All Animals? *Dev. Cell*, **22**, 903–904.
Richardson,M.K. *et al.* (1997) There is no highly conserved embryonic stage in the vertebrates: Implications for current theories of evolution and development. *Anat. Embryol. (Berl).*, **196**, 91–106.
Sander,K. (1983) The evolution of patterning mechanisms gleanings from insect embryogenesis and spermatogenesis Cambridge University Press.
Steven Poe and Marvalee H. Wake (2004) Quantitative Tests of General Models for the Evolution of Development. *Am. Nat.*, **164**, 415–422.
Stuart,J.M. *et al.* (2003) A gene-coexpression network for global discovery of conserved genetic modules. *Science*, **302**, 249–255.
Sulston,J.E. *et al.* (1983) The embryonic cell lineage of the nematode Caenorhabditis elegans. *Dev. Biol.*, **100**, 64–119.
Tadros,W. and Lipshitz,H.D. (2009) The maternal-to-zygotic transition: a play in two acts. *Development*, **136**, 3033–3042.
Yan,K.-K. *et al.* (2014) OrthoClust: an orthology-based network framework for clustering data across multiple species. *Genome Biol.*, **15**, R100.
Zhu,X. *et al.* (2007) Getting connected: Analysis and principles of biological networks. *Genes Dev.*, **21**, 1010–1024.
Cunningham, F.*, et al.* Ensembl 2015. *Nucleic Acids Res* 2015;43(Database issue):D662-669.
Domazet-Loso, T. and Tautz, D. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* 2010;468(7325):815-818.
Hopwood, N. Haeckel's embryos : images, evolution, and fraud.
Howe, K.L.*, et al.* WormBase 2016: expanding to enable helminth genomic research. *Nucleic Acids Res* 2016;44(D1):D774-780.
Piasecka, B.*, et al.* The hourglass and the early conservation models--co-existing patterns of developmental constraints in vertebrates. *PLoS Genet* 2013;9(4):e1003476.