**The Comparative ENCODE RNA Resource Reveals Conserved Features of Transcription**

in it, the ENCODE and modENCODE consortia have generated large amounts of matched RNA-sequencing data for human, worm, and fly. Uniform processing and comprehensive annotation of these data allow comparison across highly divergent animals, revealing highly conserved features of transcription. This is very different from previous transcriptome comparisons focusing on more closely related organisms \cite{22012392,23258891,23258890,22560298,21150996,20969771,22206443}. In particular, by clustering expression profiles, we discovered conserved co-expression modules shared across phyla, many of which are enriched in developmental genes. We used these to align the stages in worm and fly development, finding the expected embryo-to-embryo and larvae-to-larvae pairings in addition to a novel pairing between worm embryo and fly pupae. Furthermore, we found the extent of non-canonical non-coding transcription is similar between the organisms, per base-pair. Finally, we found the gene expression levels in all three organisms, both coding and non-coding, can be consistently predicted by their upstream histone marks using a common “universal model”.

The comparison here is based on the ENCODE-modENCODE RNA resource, available online. This comprises: (1) deeply sequenced RNA-Seq data from many distinct samples from all three organisms; (2) comprehensive annotation of transcribed elements and (3) processed, standardized analysis files, focusing on non-coding transcription and expression patterns. Where practical, these datasets match comparable samples across organisms and to other types of functional genomics data (Fig. XXX). The datasets were collected using similar sequencing technologies and uniformly processed, and they sample conditions broadly, providing a comprehensive representation of transcribed features. The resource encompasses many RNA types, including poly(A)+, poly(A)- and ribosomal-RNA-depleted RNA and short and long RNA. It represents 575 different experiments containing >67B reads. The RNA-Seq data enables a comprehensive annotation of transcribed elements, representing a capstone for the decade-long annotation efforts in human, worm, and fly. The RNA-Seq data also greatly expands the known splicing events. We find the proportion of the different types of alternative splicing (e.g., exon skipping or intron retention) is similar across the three organisms; however, skipped exons predominate in human while retained introns are most common in fly The new annotation sets have similar numbers, sizes and families of protein-coding genes to previous compilations, but the number of pseudogenes and annotated ncRNAs differ (Fig. 2, ED1). Moreover, a considerable fraction of the transcriptome results from genomic regions not associated with these annotations, representing "non-canonical transcription” (Table S3) \cite{22955620,17567993}. To characterize it, we uniformly processed the reads mapping outside protein-coding genes, pseudogenes and annotated ncRNAs to identify connected clusters (transcriptionally active regions, TARs), using a minimum-run/maximum-gap algorithm. We found that consistently one third of the bases in the genome gives rise to "non-canonical" transcription. A natural question is to what extent this transcription represents an expansion of the current established classes of ncRNAs. To address this, we identified the TARs most similar to known annotated ncRNAs using a supervised classifier \cite{21177971} (Fig. S3, Table S3). We validated its predictions using RT-PCR, demonstrating high accuracy. Overall, the predictions are only a small fraction of all TARs, suggesting that most TARs have features distinct from annotated ncRNAs and that the majority of ncRNAs of established classes have already been identified. To shed further light on the possible roles of TARs we intersected them with enhancers and HOT regions \cite{mod3,21177976,21177974,22950945,mod2,22955620}, finding statistically significant overlaps (Fig. ED5, Table S3).

Given the uniformly processed nature of the dataset and annotation, we were able to compare consistently across organisms and integrate across data types. First, we built co-expression modules (Fig. 3), extending earlier analysis\cite{12934013}. To detect them consistently across the three species, we combined across-species orthology and within-species coexpression relationships. We then searched for dense subgraphs (modules) in the resulting multilayer network, using simulated annealing\cite{17813860,15601068}. Some modules are dominated by one species, whereas others contain genes from two or three. As expected, the modules with genes from multiple species are enriched in orthologs. A phylogenetic analysis shows that the genes in the modules are more preserved across 56 animal species (Fig XXX, S2). To focus on the cross-species conserved functions, we restricted the clustering to orthologs, arriving at 16 conserved modules, which are enriched in a variety of functions, ranging from morphogenesis to chromatin remodeling (Fig. 3a, Table S2). Finally, we annotated many TARs based on correlating their expression profiles with the modules (Fig. ED5).

In addition to clustering genes, we used expression profiles of orthologous genes to align the developmental stages in worm and fly (Fig. 3b, ED4). Specifically, for every developmental stage, we identified stage-associated genes, i.e. genes highly expressed at a particular stage but not across all stages. We then counted the number of orthologous pairs amongst these stage-associated genes for each possible worm-and-fly stage correspondence, aligning stages by the significance of the overlap. Strikingly, worm stages map to two sets of fly stages (Fig. 3b). First, they match in the expected one-to-one fashion to the fly (i.e. embryos-to-embryos, larvae-to-larvae, etc). However, worm late embryonic stages also match to fly pupal stages, suggesting a shared expression program between embryogenesis and metamorphosis. The ~50 genes involved in this dual-stage mapping are enriched in functions such as ion transport and cation-channel activity (Table S2).

To get further insight into the stage alignment, we examined our 16 conserved modules in terms of the "hourglass hypothesis", i.e., all organisms go through a particular stage in embryonic development (the tight point of the hourglass or "phylotypic" stage) during which the expression divergence across species for orthologous genes is smallest\cite{21150996,22560298,21150997}. For genes in 12 of the modules, we observed canonical hourglass behavior, i.e. in the fly developmental time-course "inter-organism" expression divergence across closely related fly species is minimal\cite{21150996}(Fig. S2). Moreover, we find subset of TARs exhibit also exhibiting "hourglass" behavior (Fig. S3). Beyond looking at *inter*-organism divergence, we also investigated the *intra*-organism divergence within just *D. melanogaster* and *C. elegans*. Strikingly, we observed that divergence of gene expression across modules is minimized during the worm-and-fly phylotypic stages (Fig. 3c, Suppl.). This suggests, for an individual species, the expression patterns of different modules are most tightly coordinated (low divergence) during the phylotypic stage, but each module has its own signature (high) before and after this. One can, in fact, directly see this coordination as a local maximum in the between-module correlation for the worm (Fig. ED3). Finally, using genes from just the 12 "hourglass modules," we found that the alignment between worm-and-fly stages becomes stronger (Fig. 3b, S2). The alignment shows a gap, perfectly matching the phylotypic stage.

Next, we investigated the degree gene expression can be predicted from upstream factor-binding and chromatin-modification data and how consistent this prediction is across organisms (Fig. 4). Overall, we found consistent correlations in each of the three species between the expression level of the downstream gene and the signal of histone modifications. Around the TSS, H3K4me1, H3K4me2, H3K4me3 and H3K27ac are positively correlated, whereas H3K27me3 is negatively correlated (Figs. 4a, ED6, S4). We then integrated the signal from these histone modifications into a statistical model, obtaining high accuracy in predicting expression for both protein-coding genes and ncRNAs. The promoter-associated marks, H3K4me2 and H3K4me3, consistently have the highest contribution to the model.

A similar statistical analysis using TFs showed the correlation between gene expression and TF binding to be the greatest in magnitude right at the TSS, positively for activators and negatively for repressors (Fig. 4d). Integrated models using the TFs in each organism also achieved high accuracy for protein coding genes and and ncRNAs, with only a few TFs necessary in each of the models (Fig. ED7). In particular, models with as few as five TFs give accurate predictions. This presumably reflects an intricate, correlated structure to regulation. The relative importance of the upstream regions is more peaked for the TF models than for the histone models, likely reflecting the fact that histone modifications are spread over broader regions, including the gene body, whereas most TFs bind near the promoter.

Next, we constructed a "universal model," containing a single set of organism-independent parameters. This achieved accuracy comparable to the organism-specific models. In the universal model, the consistently important promoter-associated marks such as H3K4me2 and H3K4me3 are weighted most highly. In contrast, the enhancer mark H3K4me1 is down-weighted, perhaps reflecting the fact that signals for most human enhancers are not near the TSS. The same universal model also can predict ncRNA expression, i.e. using the same set of organism-independent parameters derived from training on protein-coding genes.

Overall, we compare the transcriptomes of three highly dissimilar metazoans, a comparison not previously attempted. Our comparison highlights fundamental principles conserved across metazoans. First, there are ancient co-expression modules, corresponding across organisms, many of which are enrich for developmentally important hourglass genes. These conserved modules have highly coordinated intra-organism expression, during the phylotypic stage, but display diversified expression before and after. We were also able to use the expression clustering to align developmental stages between worm and fly, revealing shared expression programs between embryogenesis and metamorphosis. Next, we were able to show that the transcription in all three organisms could be predicted consistently from upstream histone marks using a single set of parameters for both protein-coding genes and ncRNAs. Overall, our results underscore the importance of comparing two divergent model organisms to human to highlight conserved biological principles (and dis-entangle them from lineage-specific adaptations).

**Fig 3 - Expression Clustering**

(A) Left: Human, worm, and fly gene-gene co-association matrix; darker blocks reflect increased likelihood pairs of genes are assigned to the same module. Blocks along the diagonal represent groups of human, worm, and fly genes. Blocks from different species with off diagonal matches form cross-species modules; whereas blocks without any off diagonal matches form species-specific modules. Right: The functional enrichment of genes within each module is shown. (B) Alignment of worm and fly developmental stages based on all worm-fly orthologs. Inset shows worm-fly stage alignment using only hourglass genes is more significant and exhibits a gap (brown) that matches the phylotypic stage. (C) Expression of 16 conserved modules shows smallest intra-organism divergence during the phylotypic stage (brown).

## Fig 4 – Histone Models for Gene Expression

(A) Binding/expression correlations of two representative histone marks. (B) Relative importance of histone marks in organism-specific models and the universal model. (E) Cross-organism prediction accuracy of the histone marks- and universal-model.