DO NOT SHARE

(FOR NOW, “FOR YOUR EYES ONLY!”)

Document Conventions at<http://goo.gl/kSrS9>

===========================================

# Comparison of 3 metazoan transcriptomes

# \* Abstract

The ENCODE and modENCODE consortia have generated large amounts of RNA-sequencing data and annotation to comprehensively describe the transcriptomes of three highly divergent animals: human, worm (*C. elegan*s), and fly (*D. melanogaster*). These three organisms have fairly consistent overall statistics related to gene structure and alternative splicing. Clustering gene expression levels reveals a number of conserved modules shared between the organisms that are enriched in developmental genes. Moreover these conserved modules can be used to align the stages in worm and fly development, finding the normal embryo-to-embryo and larvae-to-larvae pairings in addition to a novel pairing between worm embryo and fly pupae. We find the gene expression levels in the organisms can also be predicted quite consistently based on their upstream histone marks or, surprisingly, just a few of their upstream-binding transcription factors. In fact, for histone marks a "universal model" with a single set of cross-organism parameters can predict expression level for both protein coding genes and ncRNAs. A principal difference between the organisms is in their pseudogene repertoire, reflecting their different histories. The human has many more pseudogenes, most of which are recently processed and stem from highly expressed genes (e.g. metabolic and ribosomal proteins). In contrast, the worm and fly have fewer pseudogenes, often from highly duplicated, environmentally responsive gene families. However for the pseudogenes in all the organisms there is consistent fraction with residual biochemical (transcription and upstream binding) activity.

===========================================

# \* Introduction

An understanding of the transcriptome is fundamental to genome annotation and to interpreting the binding events and epigenetic modifications to the genome. Furthermore, changes in the transcriptome are essential to interpreting the differences between cell types.

Over the past decade there has been an explosion in information related to the transcriptome. ENCODE (Encyclopedia of DNA Elements) and modENCODE (Model Organism Encyclopedia of DNA Elements) are two consortia that have recently published individual-organism transcriptome analyses focused on the human, worm (*C. elegans*), and fly (*D. melanogaster*) \cite{22955616,21177976,21179090,21177974}. Here, we compare the transcriptomes of these three organisms using comprehensive and matched data generated by the two consortia. In parallel, similar analyses comparing the chromatin organization \cite{mod2} and regulatory programs \cite{mod3} across human, worm, and fly are presented. These three phylogenetically distant metazoans are separated by >500 million years from a common ancestor. Since direct comparison using synteny is not possible at this distance, comparisons are made in terms of distant orthologs and overall principles of transcription. In particular, we investigate:

*PROTEIN CODING GENES & SPLICING*. We compare protein coding gene annotation across the three organisms. In relation to this, we have uniformly annotated and quantitated alternative splicing. Alternative splicing greatly expands the number of potential protein products encoded in genomes \cite{20110989,21177976,21179090,22955616,21177968,18978772,18978789}, and this phenomenon has evolved rapidly in the vertebrate lineage perhaps contributing to greater anatomic and functional complexity \cite{23258891,23258890}. We find that the relative prevalence of the types of splicing events is broadly consistent between the three organisms.

*PSEUDOGENES*. Often referred to as “genomic fossils” \cite{17568002,16574694}, pseudogenes are disabled copies of a parent protein-coding gene. However some can be transcribed \cite{22951037,17382428}, perhaps playing regulatory roles \cite{20577206,21816204}. Previously, pseudogenes have been characterized within individual genomes \cite{17099229,22951037,11160906,12560500,15860774,12083509,16925835}. Here, based on uniform cross-organism annotation, we show that the pseudogene repertoire is quite different between human, worm, and fly and that this has important implications for genome evolution. We also highlight a similar fraction of residual, pseudogenic biochemical activity in all three organisms.

*NCRNAs & NON-CANONICAL TRANSCRIPTION*. A significant portion of the non-genic regions in all three species have been found to be transcribed at a low level. The biological relevance of this "non-canonical" transcription (outside of annotated protein-coding or ncRNA loci) is debated \cite{15539566,15998911,20502517,21765801}. Here, we investigate the amount of this transcription and find a small but roughly comparable fraction in the three organisms.

*CO-EXPRESSION MODULES*. Clustering analysis on expression profiles has been useful in identifying functional modules of genes with an individual organism \cite{9843981,19114008,17021160,21177976}. Here, we develop a novel way of carrying out clustering in a coordinated fashion across species. We find that the resultant clusters fall into two categories: (1) highly conserved modules, associated with development, and (2) species-specific modules. Furthermore, we find that expression clustering, particularly based on the conserved modules, is able to "align" developmental stages in worm and fly.

*EXPRESSION MODELING*. Previous studies have shown that upstream transcription factor binding and histone modification signals are statistically predictive of gene expression levels \cite{22955978,22955616,22950368,21926158,21324173,20133639,19995984}. However, the generality of this result is unclear, i.e. does it represent ancient relationships or does each histone mark function in an organism specific fashion? Here, we show that the histone modifications and transcription factor binding upstream of each gene are comparably predictive of the level of gene expression in human, worm, and fly. Moreover, we have constructed a universal model (comprising a small number of cross-organism parameters) that works for both protein coding genes as well as ncRNAs in all three organisms.

# \* Transcriptome Data and Annotation

## \*\* Functional Genomics Data Sets Used

We have generated a wide range of data sets to sample the organisms’ transcriptomes broadly, so as to most comprehensively annotate all the transcribed products and estimate the changes in their expression (Fig 1; see Suppl). Briefly, the modENCODE data for worm and fly are derived from multiple time-points in the developmental life cycle, different environmental conditions and from whole organisms, various tissue preparations, and cell lines. The human ENCODE data sets were produced from cell lines including those derived from primary tissues and embryonic stem cells. In addition, for a number of selected analyses, we have made use of the BodyMap 2.0 data, consisting of gene expression from 16 human tissues \cite{bodymap}. The entire data set encompasses many RNA types, including poly(A)+, poly(A)- and total RNA; both short (<50 nt) and long RNA (>200 nt); and CAGE data. In this study, for the most uniform comparison of the transcriptomes of the three organisms, we focus on whole-cell poly(A)+ RNA-Seq data. This “matched-compendium” includes 571 data sets with >20 billion total aligned reads in human, worm, and fly. This extremely large collection of data sets provides power to detect many different classes of transcribed elements, including ncRNAs (see Suppl.). Furthermore, we selectively integrate the matched compendium with datasets of other RNA types.

In addition to the transcriptome data, incorporating histone modification data and transcription-factor binding data allows for a more comprehensive interpretation of transcription. A large number of such data sets have been generated for human, worm, and fly under matched conditions to those used for transcriptome profiling \cite{mod2,mod3} and we integrate these in our analyses. For embryonic tissue, in particular, many matched factor-binding and chromatin data sets are available, so these were a focus for integration. (For worm and fly, the embryo data are derived from whole embryos and their tissues, and for human, from H1-hESC, a stem cell line.)

## \*\* Annotation Sets Used

For the last half decade the ENCODE and modENCODE projects have supported annotation efforts in the human, worm, and fly genomes. To annotate protein coding and non-coding regions, these efforts draw upon a variety of evidence sources (e.g. RNA-seq data and ESTs) and synthesize the results of computational pipelines with a substantial amount of manual curation. For the comparison here, we used GENCODE v10, the output of a subproject of ENCODE, for human. For worm and fly we developed annotation sets that are based on the full RNA-Seq data sets produced by the modENCODE consortium as well as other available data. These annotations were related to WormBase WS220 and FlyBase 5.45 but extend beyond them.

The annotation of protein coding genes is fairly stable, whereas the annotation of non-coding elements is more fluid. In particular, with each iteration of the annotation, the overall number of protein coding genes has not substantially changed. For instance, for the human, the number of protein coding genes was 19,599 in the first version of GENCODE and five years later is now 20,007 (Fig S1). In contrast, the numbers of non-coding loci in the corresponding gene-sets have increased considerably. Given this fact, we have taken a two-pronged approach to comparing across organisms. For protein coding genes, we have used the current annotations and compared them without modification. However, for non-coding and pseudogene loci, we have uniformly reprocessed and harmonized the annotations to make them more comparable between organisms. Finally, to facilitate direct comparisons between the three species, we utilized sets of protein-coding gene orthologs with varying stringency in our analyses \cite{mod14} (see Suppl.). These include triplets of 1-to-1-to-1 orthologs as well as one-to-many and many-to-many orthology relationships.

# \* Comparison of Protein-Coding Genes

## \*\* Overall Characteristics

We began our comparisons of by examining the basic characteristics of the protein-coding repertoire (Fig 2). Human and worm have ~20,000 genes whereas fly has about 2/3 of that number. Fly and worm genes span similar genomic lengths but human genes span larger regions (Fig 2B); most of this difference is due to the greater size of human introns, but longer human CDSs also contribute. Individual exon lengths have similar distributions between human and worm with fly having some larger exons. Human differs most obviously from fly and worm, however, in the number of exons per gene. This is also reflected in the human genome containing more than twice as many coding bases and 5’ UTR sequence as worm despite having the same number of genes.

Given the difference in CDS size and exon complexity between the organisms, we revisited the question of domain complexity raised in the initial genome papers \cite{11237011,11181995}. We found that <10% of the protein domains found in fly and worm were specific to each phyla, whereas nearly 20% of human protein domains were not found in the two models (see suppl., Fig 2C, S2 and Associated Data File). While overall most domains are shared, we found that unique combinations of domains have substantially increased in human, with only 32% of human domain combinations shared by the other two organisms (Fig 2C), whereas the corresponding numbers for fly and worm are 64% and 54%.

In analyzing the RNA-Seq reads, we found that overall the distribution of reads is quite similar in fly and worm with ~95% of the reads mapping to annotated protein coding genes (CDSs + UTRs). The majority of reads in human also derive from protein-coding genes, but the somewhat smaller fraction (78%) reflects the fact that there is much more sequence in human outside of protein coding segments as compared to worm and fly. The RNA-Seq data also reveals the fraction of genes broadly expressed across cell lines, tissues and developmental stages. Overall, we identified 6,912 (38%), 5,180 (25%) and 5,288 (38%) broadly expressed protein coding genes in human, worm, and fly respectively (Table S1, Suppl. and Associated Data File). Across all three organisms, these genes tend to be more highly expressed and to have characteristic housekeeping functions, such as RNA processing, protein transport, and protein localization (see Table S1 and Fig S2).

## \*\* Comparison of Splicing of Orthologs

We next compared alternative splicing between the three organisms. Our short RNA-Seq reads can define individual splice junctions and exons, but do not allow unambiguous definitions across whole transcripts. Accordingly, we focus our analysis on junctions and exons. Overall, we do not find a conservation of splicing between orthologs, with no examples of orthologs preserving complete splicing structures across the three organisms. Moreover, there are only a few preserved splice sites in all three organisms (573 out of 37,517 unique splices in orthologs) (Table S2). Fig 3A depicts an example, the orthologous genes KCNMA1 (human), slo-1 (worm), and slo (fly). The exon/intron organization of the genes differs significantly between the three species, and there are no orthologous alternative exons.

## \*\* Comparison of Splicing in Annotation Sets

Using the protein-coding gene annotation sets described above we compared the number of annotated isoforms per gene in the three organisms (Fig 3B). (Note, by design the three annotation sets used give a conservative estimate of the number of alternative splicing events. See Suppl.) Consistent with recent reports \cite{18978772,18978789} we find that ~81% of human protein-coding genes express multiple mRNA isoforms. The corresponding fraction is similar in worm (~74%), but less in fly (~47%). However, fly has the greatest number of "outlier" genes that express a large diversity of isoforms; in particular, it has 100 genes with >50 isoforms. An extreme case is Dscam, which has >38,000 potential isoforms \cite{10892653}.

Overall, human has the greatest number of alternative splicing events (88,492), worm has an intermediate number (30,625) and fly has the lowest number (25,756). There are many classes of alternative splicing (e.g., exon skipping, intron retention, and alternative 5' splice sites). The proportions of the different classes are broadly similar between the three organisms (Table S2 and Fig 3 and S2). However, skipped exons are most abundant in human while retained introns are most abundant in fly (Fig 3C). This may relate to differences in the splicing mechanisms: In human, with large introns, exon definition predominates, and the splicing machinery recognizes splice sites on either side of an exon \cite{7852296}. In fly, intron definition predominates, in which the splicing machinery coordinately assembles on the splice sites on either side of an intron \cite{8164690}.

Another difference between the three species relates to mutually exclusive splicing. In fly, there are 12 mutually exclusive splicing events involving more than two exons. These range from clusters of three alternative exons to as many as 48 alternative exons in *Dscam*. In contrast, all mutually exclusive splicing events in human and worm involve only two exons. The mechanisms required to faithfully splice clusters of mutually exclusive exons containing only two and more than two exons are distinct \cite{18380340}, suggesting another fundamental difference in the splicing machinery of fly compared to that of human and worm.

## \*\* Comparison of Splicing Quantification

We next investigated how alternative splicing varies quantitatively between samples (distinct conditions, tissues or developmental stages) and organisms. Because of the limitations of short-read RNA-Seq noted above, we have only quantified individual splicing events consisting of an alternatively spliced exon and its immediate flanking constitutive exons. Furthermore, we only analyzed unambiguous events in which other annotated transcripts could not confound the interpretation and the events involve only two possible alternative splice-forms. To facilitate a consistent quantification, we uniformly processed the matched compendium with a splicing focused pipeline (see Suppl.), calculating the percent inclusion of the alternative exon in each sample. The majority of events in each species are either primarily included or skipped (Fig 3D). However, when examining each individual event class, there are some distinctions. For example, most skipped exon events in each organism have one splice-form that greatly predominates, though the nature of that differs between organisms. In particular, skipped exons are most often absent in fly and present in worm (Fig 3D). For tandem UTR events, human has a striking peak at an inclusion value of ~50% indicating that the long and short forms of the UTRs are expressed at equal levels on average.

We next examined the dynamics of alternative splicing for each splicing event by calculating the maximal change in percent inclusion over all pairs of samples for each species. We call this quantity the "switch score" and binned the splicing events into those that varied strongly (switch score >50%), moderately (25-50%), or weakly (0-25%). The majority of splicing events in fly and worm change dramatically between samples (Fig 3E).

Finally, we examined the conservation of the sequences associated with splicing, specifically skipped exons and their flanking introns. In all three species, the strongly varying exons (those associated with the highest switch scores) and the adjacent portions of their flanking introns are more conserved than moderately varying exons, and the exons that vary the least are also the least conserved (Fig 3F).

# \* Pseudogenes

One particular type of non-coding element that is directly related to protein coding genes is the pseudogene. We uniformly annotated the pseudogenes in all three organisms using a combination of automated pipelines and manual curation (see Suppl.). We then compared the pseudogenes on a number of levels ranging from biotype distribution to transcription.

## \*\* Large Differences in Pseudogene Complements Reflect Evolutionary History

Overall, the pseudogenes differ greatly between the three organisms, reflecting the unique evolutionary history of each (Fig 4). The human genome has ~12-fold more pseudogenes than worm, which has ~8-fold more than fly; ratios that do not match their relative genome sizes or gene counts. Based on their mechanism of formation (Fig 4A, Table S3a) \cite{12034841}, pseudogenes can be classified into two groups, duplicated and processed, with the later resulting from retrotransposition. We found that processed pseudogenes are enriched in human, whereas there are more duplicated pseudogenes in worm and fly. The later enrichment can be related to the relatively high gene duplication rate in worm and fly \cite{11861885,11230161,21295484}. Moreover, previous studies \cite{12572619,1806330,9402741} suggest that the scarcity of fly pseudogenes can be explained by the high rate of DNA loss inherent to the fly genome \cite{12572619,9501496}.

Using sequence similarity to parent genes, we inferred the time of origin of pseudogenes. Most informative is the fraction processed of the total number of pseudogenes at different ages (see Fig 4B and Table S3). In humans, a prominent peak of processed fraction at 93% similarity corresponds to burst of retrotransposition ~40 million years ago, at the dawn of the primate lineage when the bulk of human pseudogenes were created. By contrast, in worm, older pseudogenes (~45% similarity) tend to be processed, whereas younger ones are more likely to be duplicated. This preponderance of recently duplicated pseudogenes in worm relative to human, might relate to large block duplication events in the recent evolutionary history of the worm \cite{19622155,19289596,11230161,11861885}. Moreover, the majority of worm pseudogenes are near the telomeres (Fig. 4C), a location characterized by a high number of recombination events and rapid gene evolution; in contrast, human pseudogenes (both duplicated and processed) are evenly distributed along the length of the chromosome.

We find that pseudogenes arise from different progenitors in the three organisms. In particular, by analyzing pseudogenes associated with parents that are 1-1-1 orthologs, we observe that there is no similarity in the pseudogene complement of orthologous genes. In fact, not one of the triplets of 1-1-1 orthologous genes have associated pseudogenes in all three species (see Associated Data Files, Fig S3). An example is shown in Fig 4D: the number of RpS6 pseudogenes varies significantly among the analyzed genomes, with human having 25 mostly processed pseudogenes spread randomly over the whole genome, fly having three duplicated pseudogenes clustered near the RpS6 gene and worm having no RpS6 pseudogenes at all.

We next generalized this analysis from 1-1-1 orthologs to parent gene families (see Suppl. and Fig. S3). Fig 4E shows that different families dominate the pseudogene repertoire in each of the organisms, with only three families being amongst the largest in all three -- kinase, histone and P-loop NTPase. For instance, motor protein genes dominate fly. Worm tends to be dominated by chemoreceptor pseudogenes, perhaps reflecting the many duplications of this family in nematode evolution \cite{19289596,18837995} and the fact that this family is rapidly evolving \cite{11961106}. Human also has a 7-TM family as a top pseudogene family, reflecting the duplication and divergence of the olfactory receptors. However, the human pseudogene set tends to be dominated overall by ribosomal proteins and of metabolic enzymes; these genes are highly expressed, making them likely targets of retrotransposition \cite{16504170}. Specifically, the distribution of ribosomal pseudogenes reflects the general burst of retrotranspositional events 40 million years ago (Fig 4B), and these pseudogenes tend to be more recent than those of olfactory receptor genes (Fig S3).

## \*\* Pseudogene Activity

In addition to examining their evolution, we looked for signs of biochemical activity for each pseudogene. First, we computed an expression value based on RNA-Seq data and obtained 1,441, 143, and 23 potentially transcribed pseudogenes in human, worm, and fly respectively (see Suppl., Fig 4A). This represents a fairly consistent fraction of the total pseudogene complement in each organism. Interestingly, we found a subset of these (~13% in human and ~30% in worm and fly) that have discordant transcription patterns with their parent genes over multiple samples (Fig S3). The results also indicate pseudogenes are less broadly transcribed than protein coding genes. Specifically, only 5.1%, 0.69%, and 4.6% are broadly expressed in human, worm, and fly, respectively (see suppl. and table S1). Moreover, a substantial number of pseudogenes are expressed in only a single cell line or developmental stage (Fig S3). Finally, we found that the parent genes of broadly expressed pseudogenes tend to be broadly expressed as well (Fig S3).

There are a number of additional indicators of pseudogene biochemical activity, including the presence of active TF and RNA Polymerase II (Pol II) binding sites in their upstream regions of and proximal regions of "open chromatin" (as determined from histone modification data). We thus integrated the transcriptional information with these other data to create a comprehensive map of pseudogene activity (Associated Data Files, Fig 4A and S3), grouping pseudogenes into different categories. At one extreme, completely "dead" pseudogenes -- not transcribed, lacking any evidence of TF and Pol II binding and active chromatin marks -- represent ~20% of the total pseudogenes in each of the three organisms. On the other extreme, there are very few pseudogenes (98 in human, 40 in worm, 6 in fly) that are both transcribed and simultaneously exhibit all other activity features (namely open chromatin, transcription factor and Pol II binding), despite the presence of mutations that disrupt the protein coding sequence. We label these pseudogenes as “highly active”. The majority of pseudogenes are intermediate between these two. (see Fig 4A). Such pseudogenes have only a few of the classic indicators of activity.

# \* ncRNAs & Non-Canonical Transcription

Unlike pseudogenes, ncRNAs are not defined in terms of proteins. Here we divide the potential ncRNAs produced by the genome into two groups: (1) those that correspond to well-known and annotated classes of ncRNAs (e.g. micro-RNAs) or (2) those that arise from regions previously un-annotated as producing mature transcripts (“non-canonical transcription”) \cite{22955620,17567993}. To consistently characterize the first group across the three organisms, we took the perspective that, unlike protein coding genes, only a subset of ncRNAs, are annotated consistently across organisms. Thus we first built a subset of the annotations ("the comparable ncRNAs") that is directly comparable between the organisms. Next we added onto this set other ncRNAs that are well defined but inconsistently described between organisms. Finally, after removing all of these from the genome, as well as the exons from protein coding genes and pseudogenes, we have the fraction of the genome that is effectively un-annotated with respect to transcription, except for potentially being part of an intron. To find out how much transcription exists in this fraction we subjected each genome to uniform computational pipelines.

## \*\* Uniform Comparison of Annotated ncRNAs

The ncRNAs consistently annotated between the three organisms are shown in Fig 5C (more details on this subset in the Suppl.). In particular, there are comparable numbers of tRNAs in humans and worms with about half as many in fly. The number of lncRNAs in human is considerably greater than in worms and flies but the genomic coverage percentage in the three species is, in fact, similar (0.37%, 0.18% and 0.68% for human, worm, and fly). Finally, humans have considerably more pre-miRNAs, snoRNAs and snRNAs compared to either worm or fly. Pre-miRNAs are cleaved from longer primary-miRNA (pri-miRNA) transcripts, which are either part of protein-coding genes (where the pre-miRNAs are spliced out of introns) or comprise distinct intergenic non-coding loci. Pri-miRNA transcripts are incompletely annotated at present; however, comparison of a small but similar number of examples indicates that intergenic pri-miRNA transcripts are substantially shorter in worm (~0.37kb) as compared to fly (7.4kb) or human (20kb).

Overall, the annotated ncRNAs comprise a smaller fraction of genomic coverage and RNA abundance (fraction of reads) relative to protein coding genes (Fig 5A, Table S4); the annotated ncRNAs (dominated by miRNAs and lncRNAs) are also less broadly transcribed (only 4.4%, 3.6% and 7.8% in human, worm, and fly respectively, Table S4).

The well-known ncRNA classes that are not comparably annotated include ribosomal RNAs, which are inconsistently represented in the underlying genome sequence due to their repetitive structure (see Suppl.), and piRNA precursors, which appear to be fundamentally different in the three organisms. In particular, comparison of piRNA precursors reveals a small number of well-annotated loci in fly and human \cite{16751777,16751776,17346786}, which can, nevertheless, occupy large genomic spaces (many >50kb). In contrast, each worm 21U locus generates 26 nt transcripts that are processed into 21 nt products, with >35,000 21U genes recognized \cite{23260138}. Interestingly, while 21U genes are very short in worm, their aggregate genomic length is within ~2-fold of the well-annotated intergenic piRNA clusters in fly and human.

## \*\* Uniform Quantification of Non-canonical Transcription

After removing all annotated ncRNAs as well as exons of protein coding genes and pseudogenes, we are left with regions of each genome that that are not transcribed into annotated, processed RNAs. A considerable number of reads map into these "unannotated" regions (Table S4), and we uniformly processed them to identify transcriptionally active regions (TARs) (connected clusters of reads, also known as transcribed fragments or “transfrags” \cite{15539566,15998911}), using a minimum-run/maximum-gap algorithm with consistent parameters chosen for the three organisms (Fig 5C, Suppl.). We found that a significant portion of all three genomes gives rise to detectable "non-canonical" transcription; roughly an additional one-third (32 to 37%) of each genome is transcribed at above threshold levels (see Fig. 5A). Much of this occurs in the introns of annotated genes; the remaining transcription detected (201 Mb, 16 Mb, and 14 Mb in human, worm, and fly) is intergenic and occurs at low levels, comparable to the levels of transcription detected in introns in the matched compendium (see Table S4). In addition, only ~1% of the TARs are found to be broadly expressed in each of the three organisms (see Table S1). Overall, the total fraction of the genome that is transcribed -- including intronic, exonic, and non-canonical transcription -- is consistent with that previously reported for human in Djebali et al. \cite{22955620} despite the methodological differences in the analysis pipelines (see Suppl. and Fig. S4).

Finally, we identified the subset of the TARs that are most similar to existing annotated ncRNAs by applying a supervised machine-learning approach \cite{21177971} (see Suppl.). Our machine-learning classifier integrates expression from multiple RNA-Seq experiments (beyond those in the matched compendium) with other features such as RNA secondary structure, sequence conservation, chromatin modification and transcription factor binding. It is trained on a gold-standard set of annotated ncRNAs (essentially the comparable ncRNAs, above), producing predictions that are conservative extrapolations from these. In worm, many of the supervised ncRNA predictions have been previously validated \cite{21177976,21177971}. Here, we carried out analogous validations in fly and human using RT-PCR (see Suppl.). Moreover, further validations of similar ncRNAs predictions have also been carried out independently in these organisms as well \cite{16951679,22955620,23104886}. Overall, the number of supervised ncRNA predictions is only a small fraction of the set of all TARs. Since these represent a fairly conservative set of novel ncRNAs that appear similar to the annotated ncRNAs, it may be that the majority of ncRNAs similar to those annotated have been identified

## \*\* Analysis of Antisense Transcription

So far, the detected non-coding transcription has, by definition, been restricted to avoid overlapping annotated transcription on the opposite strand. While a fully strand-specific analysis is not possible across all three species, we can identify orthologous loci that exhibit conserved antisense transcription, including the well-studied mammalian loci *Dicer-1* and *CTCF*. While the protein coding genes tend to be well-conserved across diverse phyla, "positionally equivalent" (relative to positions in an ortholog gene) antisense transcripts are often poorly conserved even between closely related mammals \cite{16290135, 16683030}.

We identified 1,629 human, 303 fly and 18 worm protein-coding genes with annotated transcripts antisense to a mature RNA (see Associated Data Files). We find that for these human genes there is a significant enrichment of orthologs (619) with worm and fly (see Suppl.). None of these orthologs have antisense annotations that are positionally conserved across all three organisms. However, we do find that 27 of the 619 human orthologs have positionally conserved antisense annotations in fly. These represent a conservative estimate of the amount of antisense transcription as they rely only on annotation. We scanned for further cases of antisense transcription by identifying TARs in fly strand-specific RNA-Seq data \cite{mod9} (see Suppl.). We found 1,721 fly protein-coding genes that exhibited antisense transcription in a testes sample that was enriched for orthologs (127 out of 1,721) with positionally conserved transcription in human (28% enrichment).

## \*\* Relationship of HOT Regions and Enhancers to Non-canonical Transcription

TF binding sites distal to genes have been associated with RNA expression \cite{20463730}. We determined the degree to which the characterized non-canonical transcription contributes to this effect in enhancers \cite{mod2,22955620} and distal HOT (high-occupancy target) regions \cite{mod3}. HOT regions have an overrepresentation of different transcription-factor binding sites \cite{21177976,21177974,22950945} and have previously been suggested to be associated with transcription; distal HOT regions are the subset of HOT regions that are not in promoters. We overlapped both distal HOT regions and enhancers with our TARs and supervised ncRNA predictions and found a strong, statistically significant overlap in all three organisms, compared to a randomly shuffled control (see Fig 5C, Suppl. and Table S4f). This could represent "enhancer RNAs", i.e. RNAs that are independently transcribed from enhancer regions \cite{20393465}. Moreover these results are consistent to those found in \cite{mod2} based on a different analysis with GRO-seq data.

# \* Expression Clustering

## \*\* Simultaneous Clustering in the 3 Organisms

With the three main elements of the transcriptome and annotation -- genes, transcripts, pseudogenes and ncRNAs -- uniformly characterized, we are now in a position to study how these elements function together as evinced from expression correlations over our many samples. To detect co-expression modules consistently across the three species, we developed a method that combines expression correlations, which represent association between genes within an organism, and orthologous gene pairs, which represent evolutionary associations. We separately construct co-expression networks for each of the three species (see Suppl.) and then combine them via connecting orthologs from different species to form a multiplex network. We searched for densely connected modules in this multiplex network, which could be genes in a single species connected by expression edges or genes across species connected by ortholog relationships. Our algorithm is based on a q-state Potts model \cite{15601068} with an energy function that takes into account both co-expression and ortholog links (see Suppl., Fig S5b-c). The assignment of nodes to different modules is given by the minimal energy state of the system, which was obtained by simulated annealing \cite{17813860} (see Suppl.). We repeated the procedure multiple times to ensure the assignment of modules is robust (see Suppl).

The cross-species modules found include genes from human, worm, and fly; the relative abundance of genes from each species varies from module to module. As shown in Fig 6A, there are modules dominated by genes from a single species as well as modules with genes from all three. As expected, the latter have more orthologs than the former and therefore are more conserved. This conservation is further supported by a more comprehensive phylogenetic analysis in which genes from the modules were compared across 55 animal species (Fig 6A) and by analyzing the GO terms of each module for cross-species consistency (see Suppl. and Fig S5d1). After deriving our initial cross-species clustering, we refined it by performing a clustering on just orthologs. This gave rise to 16 conserved modules with at least 30 orthologous genes in each (10 orthologous triplets, see Table S5). Each of these modules is enriched in variety of distinct functions, ranging from RNA processing to protein catabolism to nervous system development (see Fig S5d2).

## \*\* Using Expression Modules to Annotate ncRNAs

Like conventional clustering analysis, our cross-species modules can be used to infer biological roles for genes (see Fig S5a). Furthermore, we assigned thousands of ncRNAs and TARs to the conserved modules by correlating and anti-correlating their expression profiles with the orthologous genes in the modules. Specifically, the 16 conserved modules cluster with 1706, 79, and 701 annotated ncRNAs and 8598, 9029, 4750 TARs in human, worm, and fly respectively (see Suppl. for details). The co-expression of these ncRNAs and TARs with orthologous genes suggests that they might play related functional roles. (We provide module annotations and associated GO terms for ncRNAs and TARs in the three organisms in the Associated Data Files.)

For instance, our clustering includes the annotated ncRNA mir-10, which regulates neighboring HOX genes in fly and human \cite{21210939}. We found that primary transcripts for mir-10 in fly and mir-10a in human both were highly correlated with orthologous HOX genes (Dfd in fly, HOXB4 in human) in the same co-expression module (Fig 6B). We found no analogous miRNA from worm in the same module with Dfd, HOXB4, and the worm ortholog lin-39, but found several novel worm TARs highly co-expressed with lin-39. In another example, we identified TARs both anti-correlated and correlated (Fig 6B and S5e) with the sarcoglycan complex subunit gene (SGCB in human, sgcb-1 in worm, and Scg in fly), which is part of a conserved module (#5) enriched for larval locomotory behavior.

## \*\* Conserved Modules Exhibit "Hourglass" Expression Patterns

We further investigated the conserved expression modules in the light of the "hourglass hypothesis" that posits that all organisms go through a particular stage in embryonic development (the tight point of the hourglass or the "phylotypic" stage) in which expression differences between orthologous genes are the smallest \cite{21150996,22560298,21150997}.

We examined the expression divergence between genes in *D. melanogaster* and their orthologs in closely related fly species. Using microarray data \cite{21150996}, we found that in 12 out of our 16 conserved modules, the expression divergence during the phylotypic stage is narrower (see Figures 6C and S5g). This is consistent with the canonical observation of the hourglass hypothesis, but we now see it in terms of our 16 modules, derived from cross-species clustering. Moreover, most strikingly, during periods consistent with the documented phylotypic stages of worm and fly \cite{21150996,22560298}, the expression of genes in our 16 modules are the most tightly coordinated with a single organism. This observation is evident from the high expression correlation observed between modules in worm, and the narrow expression difference between modules in fly (see Figure 6C). It suggests that each of the 16 modules has its own expression profile before and after the phylotypic stage whereas the modules follow a similar expression pattern during the phylotypic stage. Moreover, as the observation does not arise from cross-organism comparison but from within our RNA-Seq datasets, it opens a way to interpret the hourglass hypothesis in terms of the coupling of genes within an organism. Finally, it is worthwhile to notice that, apart from protein coding genes, ncRNAs and TARs exhibit similar hourglass behavior in their expression, highlighting a potential developmental role for these non-coding elements (see Fig S5f).

## \*\* Aligning Stages in Development with Conserved Modules

Additional insight into the relationship between the conserved modules and the hourglass hypothesis can be obtained from comparing the stages in the developmental timecourses in detail. In particular, by comparing the expression profiles of orthologous genes, we were able to align the developmental stages of worm and fly. To do this, we first identified stage-associated genes, genes highly expressed at a particular developmental stage but not always highly expressed across all stages, for every fly and worm developmental stage (see Suppl.). Then for every possible pair of fly and worm stages, the number of orthologous gene pairs amongst their stage-associated genes were counted. A statistically significant maximal overlap is used to align the stages. We found that the worm stages are mapped to two sets of fly stages (Fig 7). In the first set, worm development is matched in the expected one-to-one fashion to the fly (i.e. embryos to embryos, larvae to larvae, etc). In the second set, worm late embryonic stages are matched to the fly pupal stages, suggesting, strikingly, a shared expression program between embryogenesis and metamorphosis.

We then repeated the stage mapping analysis using just the genes from our conserved modules exhibiting the hourglass pattern. We found that the stage-mapping alignment observed above becomes stronger (exhibiting more significant P-values. See Suppl. and Fig S5i). Moreover, the alignment based on the hourglass genes shows a gap, perfectly matching the phylotypic stage (Fig 7). This gap is understandable in terms of the finding above that at the phylotypic stage the expression values of the genes in the modules converge, i.e. suggesting that none of them would be phylotypic stage specific. Finally, by excluding the conserved modules enriched in housekeeping genes, we got proportionately an even stronger signal (Fig S5i). These observations indicate the importance of a number of the conserved modules to development (in particular, Modules 2,4,5,11 and 12 from table S5).

# \* Modeling Gene Expression

Levels of expression are related to binding of TFs and the modification of histone proteins in the upstream region of genes. In this final section, we quantitatively integrate the transcriptome data with TF binding and chromatin data searching for statistical relationships between them. To allow for more precise matching across organisms and for better integration with the factor binding and histone modification data, we focus on the embryo, which has the best matched data-sets.

## \*\* RNA Polymerase II vs Gene Expression

First, to examine the relationship between polymerase binding and gene expression we plotted the values of these two quantities against each other for every gene (Fig 8A, Fig S6). Overall, we found a substantial correlation (Spearman r=0.67 in human, 0.62 in fly, and 0.64 in worm). For the majority of the orthologs (~75%) this "normal behavior" was observed consistently in all three organisms (Fig 8B), irrespective of the actual amount of expression and binding (see Suppl. and Fig S6b). However, a number of the genes deviated substantially from the trend, exhibiting in at least one organism “stalled-like” or “burst-like” behavior (high levels of Pol II binding and low gene expression or high expression and low binding).

## \*\* TFs vs Gene Expression

We next looked for correlations between gene expression and the binding of TFs (see Suppl.). For regions centered on the TSS, we calculated the correlation between the expression level of the downstream gene and the degree of upstream TF binding (in terms of average signal). Fig 8C shows some examples of these correlations. We found in general that this correlation is the greatest in magnitude, either positively for activators or negatively, for repressors at the TSS, and declines sharply away from it.

To investigate the correlations more systematically, we built a statistical model in each of the three organisms. The model integrates binding signals upstream of genes of many different TFs (see Suppl.) to predict gene expression. It shows high accuracy on both protein-coding genes and annotated ncRNAs (see Suppl. for details on the methods and ncRNAs used) (Fig 8D top). As expected, the regions around the TSS contribute most to the model, and this effect is most evident in human, perhaps reflecting its more precise TSSs. (Note, in the worm TSS definition is hampered by presence of splice leaders in worm transcripts.)

Surprisingly, only a relatively small number of TFs are necessary for predicting expression (Fig S6). In particular, while there are ~1400 total TFs in human, ~900 in worm, and ~750 in fly \cite{19274049,16420670,16613907}, most models with as few as 5 TFs make successful predictions. This presumably reflects the fact that the binding patterns of different TFs are not independent of each other; i.e. there exists an intricate inter-correlated structure to regulation, giving rise to a redundancy of most TFs for statistically predicting gene expression. In Fig 8E we compare model success across the three organisms. By examining the number of TFs required to achieve 90% prediction accuracy of a full (>30 TF) model, we find fewer TFs are needed in human than in worm or fly, perhaps reflecting that TF binding has an even more correlated structure in human (in the samples studied).

## \*\* Histone Marks vs Gene Expression

We performed an analysis similar to our "TF-model" using histone marks. Overall, these have similar correlation patterns with gene expression in all three species, exhibiting a complex spatial structure around the TSS (Fig 8C). In all three organisms, H3K4me1, H3K4me2, H3K4me3 and H3K27ac are positively correlated with gene expression, whereas H3K27me3 shows a negative correlation. In contrast, H3K36me3 shows positive correlation in worm and fly, but a weak negative correlation, particularly at the TSS, in human. (In worm and fly, high positive correlation of H3K36me3 with expression was achieved mainly in the gene body downstream of the TSS.)

We then integrated the histone marks in each organism in order to develop statistical models to predict gene expression (see Fig S6c). Overall, each model achieved high cross-validated accuracy on protein coding genes (Pearson's r of 0.81 in human, 0.73 in worm, and 0.84 in fly) (Fig 8F). The models did almost as well on ncRNAs (Fig 8D). As for the TFs, we analyzed the contribution to the model of each upstream location (Fig 8) and computed the relative importance of different histone marks to each model; similar patterns were observed in the three species (Fig 8E). For example, the promoter-associated histone marks, H3K4me2 and H3K4me3, achieved the highest relative importance in all three organisms.

Comparing the TF and histone mark models, we find the relative importance of the various upstream locations are more peaked near the TSS for the TF model, presumably reflecting the fact that histone modifications can be spread over wide regions, including the gene body, whereas functionally important TF binding is more confined to the promoter. This fact is further manifest in the relative success of both classes of models in predicting protein-coding genes vs. ncRNAs.

The histone models perform more similarly on protein-coding genes and ncRNAs than the TF-models – understandable, given the greater dependence of the TF model on the exact positioning of the TSS and the more precise TSS annotation for protein-coding genes compared to ncRNAs. Indeed, for many ncRNAs such as miRNAs the primary transcript is not annotated at all. In addition, there may be some TFs specifically regulating non-coding RNAs (e.g. analysis of the targets of GEI-11 \cite{mod3} show that it mostly binds near ncRNAs.)

## \*\* An Organism-Independent Universal Model

Both the TF and histone-mark models were constructed in an organism-specific fashion, e.g. the chromatin features in an organism were used to predict the expression in the same organism. Overall, these organism-specific models achieved the highest cross-validated accuracy (Fig 8F). Nevertheless, given the similarities of the histone models in all organisms, we tried to construct a universal, organism-independent model trained on data from all three organisms. This model (containing a single set of organism-independent parameters) predicts gene expression levels with high accuracy in all organisms, achieving accuracy comparable to the organism specific models (Fig 8F). In terms of the relative importance of each feature, we found the universal model does not attribute the marks with the same importance as the organism-specific models. As might be expected, histone marks with more consistent importance in different organisms tend to be up-weighted in the universal model, whereas less consistent marks tend to be down-weighted. As a result, promoter-associated marks like H3K4me2 and H3K4me3 achieve the highest relative importance in the universal model. In contrast, the enhancer mark H3K4me1 is down weighted since it shows much lower importance in the human model at the promoter in comparison to the worm and fly models, perhaps reflecting the fact that human enhancers are distant from TSS and thus are not captured by the H3k4me1 signal nearby the TSS. The universal model also down-weights H3K27me3, a repressive mark, consistent with the observation that repressive marks are less consistent in their behavior across the organisms than activating marks \cite{mod2}.

The universal model was also used to predict non-coding RNA expression (Fig 8F) without further alteration, i.e. the same set of organism-independent parameter derived from training on protein coding genes was applied to ncRNAs. (Note, other types of training are also possible: see Fig S6.) The model does quite well in human and fly however in worm the prediction accuracy suffers, perhaps because of less accurate TSS definition.

# \* Discussion

We compare the transcriptomes of a divergent set of animals -- human, worm, and fly -- revealing aspects of transcription conserved over long evolutionary time scales that should be fundamental to metazoans. Finding the same general results in human and model organisms provides added confidence in the robustness of the human studies and provides a rationale for detailed experimental study in the models. Our comparison, of course, also identifies differences, which can be connected to lineage-specific changes.

Starting with the commonalities, the extent of non-canonical transcription is similar in each organism, taking into account the larger size of the human genome. While this transcription represents a considerable fraction of each genome, we observe that the vast majority of these regions are transcribed at much lower levels than protein coding genes or annotated ncRNAs. Biological processes are often noisy. For example, in protein synthesis, mistakes occur about 1 in every 20,000 amino acids \cite{19129838} and DNA polymerases have been postulated to have error rates limited by the power of random genetic drift \cite{20594608}. It seems likely that transcription itself may be imprecise, resulting in some "noise". On the other hand, some *bona fide* functional transcripts may be quite rare or present in higher copy numbers but in relatively few cells, especially in multi-cellular samples. Moreover, these genomic regions may represent incidental or opportunistic transcription at locations accessible to Pol II, potentially constituting a substrate for evolution. Alternatively, they may be cell-type or condition specific transcription that we do not yet have the resolution to observe. To establish a biological role for any of these RNAs will require specific efforts.

We find that we can build statistical models that relate the properties of the 5'-ends of genes, in terms of histone modifications and transcription factor binding, to their level of gene expression. Somewhat surprisingly, it is possible to build an organism-independent model that predicts transcription in all three organisms with a single set of parameters for both protein coding genes and ncRNAs. In evolutionary terms, the high predictive power of the same set of histone marks for gene expression in all three species implies that a basal transcription machinery involving control of chromatin architecture was established at least as early as the bilaterian common ancestor and has remained relatively stable since then. Furthermore, the predictive ability of the "TF-model" with only a few TFs underscores the correlated nature of upstream binding by many TFs .

Clustering of gene expression over all samples in the matched compendium reveals a conserved set of modules consisting of protein-coding genes and ncRNAs, which carry out similar expression programs. There is suggestive evidence that many of these modules may be connected to the hourglass model of transcription in embryo development, highlighting their importance developmentally. Furthermore, expression clustering, particularly with the conserved, "hourglass" modules, can be used to align the developmental stages in the worm and fly, revealing the expected pairing (i.e. embryo-to-embryo and larvae-to-larvae) and also a novel additional one between worm embryo and fly pupae.

One big difference between the organisms is in their pseudogenes. In particular, there is very little commonality in the pseudogene repertoire across organisms, reflecting their very different histories. In particular, most human pseudogenes arose from a relatively recent burst of retrotranspositional activity. In contrast, the worm and fly pseudogenes reflect more the dead by-products of rapidly evolving and duplicating families (e.g. the chemoreceptors in the worm.) However, all these organisms have a roughly similar fraction of transcribed pseudogenes, perhaps hinting at a similar rate of decay for pseudogenic activity.

Our comparison connects the human genome to worm and fly, where powerful experimental approaches allow for functional tests. Overall, it underscores the importance of using a range of model organisms, as this enables us to disaggregate lineage-specific adaptations from conserved biological principles. We imagine this type of comparison could be extended in the future to encompass additional models.

# Figure Captions

## Fig 1 - Data Sets

Transcriptome profiling data have been generated for human, worm, and fly across a variety of tissues, cell lines, and developmental stages. The availability of these data sets for each organism is indicated by colored symbols next to the corresponding developmental stages and/or tissues. The symbol color represents organism (human: red, worm: green, fly: blue), whereas the shape represents different origins of the data set (circle: whole organism, square: tissues, triangle: cell line). Each symbol is followed by the number of data sets generated in that category. A detailed description of the complete data sets can be seen in the Suppl.

## Fig 2 - Summary Statistics for Protein-coding Genes

Summary statistics for the protein coding gene annotations. (A) (top) Number of bases in 5’ UTRs (5’), 3’ UTRs (3’), coding exons (C), and introns (I), in millions of bases (black) and as percentage of the genome (grey). (bottom) Number of mapped reads in the same categories as above, in millions of reads (black) and as percentage of the total number of mapped reads (grey). For both tables, relative fractions in each category are visualized in radar plots (right). (B) Distributions of key summary statistics; note that the x axes are in log scale. (C.) (left) Venn diagram of protein domains (from the Pfam database version 26.0, \cite{22127870}) present in annotated protein-coding genes in each species. (right) Shared domain combinations. (For more information on domain combinations, see Fig S2a.)

## Fig 3 - Alternative Splicing

(A) Orthologous genes do not share the same exon/intron structure or alternative splicing. (B) Distribution of the number of isoforms per gene. (C) Comparison of the fraction of various alternative splicing event classes in human, worm, and fly -- skipped exons “SE”, retained introns “RI”, alternative 3' splice sites “A3SS”, alternative 5' splice sites “A5SS”, alternative first exons “AFE”, alternative last exons “ALE”, tandem 3' UTRs “TandemUTR”, coordinately skipped exons “CSE”, and mutually exclusive exons “MXE”. (D) Percent inclusion for different events per species. (Selected events are shown. All events are in the Fig S2f.) (E) Switch scores per species. (See text for definition.) (F) Sequence conservation of intron-exon junctions in various switch-score groups in human, worms, and flies.

## Fig 4 - Pseudogenes

Triway comparison of pseudogenes in human, worm, and fly. (A) Pseudogene statistics. The pie charts show the pseudogene distribution as function of biotype in the three organisms. There are three classes: processed, duplicated, and ambiguous pseudogenes (the biotype could not be determined based on the available data) . The tree charts differentiate the pseudogenes in human, worm, and fly based on their transcriptional evidence and additional activity features (namely TF binding, Pol II binding, and active chromatin). The pseudogenes are classified in three groups: “highly active” (are transcribed and have all the additional activity features), “zombie” (can be either transcribed and nontranscribed and have only some of the additional activity features), and “dead” (are nontranscribed and have no additional activity features). A detailed description of the different activity classes is shown in the Fig S3. (B) Sequence analysis of pseudogenes. Step plot of the relative fraction of processed pseudogene as function of age. The pseudogene age is defined by the percentage sequence similarity to parent gene. See Fig S3 for additional information regarding the differential age binning in the three organisms. (C) Distribution of pseudogenes as function of chromosomal localization (end/telomeric site vs centre/centromeric site). The chromosome centre in human was selected as the centromere, while in worm and fly, it was defined by the geometric centre. (D) Orthology. Distribution of RpS6 pseudogenes by biotype in human, worm, and fly. The three 1-1-1 orthologous RpS6 parent genes are depicted as grey ovals. (E) Pseudogene distribution in the top protein families (PFAM) for human, worm, and fly. The full description of the family types and the PFAM IDs are shown in Fig S3.

## Fig 5 - ncRNAs and Non-canonical Transcription

Summary of the number of annotated ncRNAs, supervised ncRNA predictions and the estimates of the amount of non-canonical transcription in each of the three genomes. (A) Table of annotated ncRNAs (miRNAs, tRNAs, snRNAs, snoRNAs, lncRNAs and piRNAs). We also present estimates for the amount of non-canonical transcription detected (TARs) as well as supervised ncRNA predictions in each of the three organisms. This transcription is also subdivided into the fraction that occurs within introns of protein-coding genes as well as the fraction that overlaps transposable elements. See Table S4 for the fraction of read counts that map to each of these elements. (B) ROC-like plots for predicting the amount of non-canonical transcription using the sets of expressed annotations for each organism as a gold standard. The red (fly), green (worm), and blue (human) distributions show the exon discovery rate and novel TAR discovery rate for the full set of parameters using a minimum-run/maximum-gap/threshold algorithm. (C) Graph showing the overlap of enhancers \cite{mod2} and distal HOT regions \cite{mod3} with supervised ncRNA predictions and TARs in human, worm, and fly. We find that all overlap of enhancers and distal HOT regions with respect to both supervised ncRNA predictions as well as TARs are significantly enriched compared to a randomized expectation (see Suppl.). (D) Positionally conserved antisense transcription in human, worm, and fly. Antisense transcription at a monocarboxylic acid transporter locus in human (top) and fly (bottom). In human and fly antisense transcription appears specific largely to the testes and is due to the independent transcription of a long non-coding RNA.

## Fig 6 - Expression Clustering

(A) Gene-gene co-association matrix for human, worm, and fly genes. The matrix elements represent the frequency at which two genes are assigned to the same module in the multiple runs of the integrated clustering algorithm. Blocks along the diagonal represent modules of human, worm, and fly genes. Blocks from different species with high co-appearance frequency (off-diagonal positions) form cross-species modules. Isolated blocks form species-specific modules. (B) The left figure shows ncRNA/TARs highly correlated with corresponding HOX orthologs in worm (lin-39), fly (Dfd) and human (HOXB4). The expression of ncRNA, mir-10, correlates strongly with Dfd in fly (r=0.66, p<6e-4 in fly), and hsa-mir-10a, with HOXB4 in human (r=0.88, p<2e-9). A TAR (chrIII:8871234-2613) strongly correlates with lin-39 (r=0.91, p<4e-13). The right figure shows three TARs in human (chr19:7698570-7701990), worm (chrII:11469045-440), and fly (chr2L:2969620-772) that are negatively correlated with the expression of three orthologous genes: SGCB in human (r=-0.91, p<3e-16), sgcb-1 in worm (r=-0.86, p<2e-7), and Scgbeta in fly (r=-0.82, p<4e-8). (C) In the top panel, the expression levels of a conserved module in *D. melanogaster* and its orthologous counterparts in other 5 Drosophila species are plotted against time. The x-axis represents the middle time points of two-hour periods at fly embryo stages. The boxes represent the log10 modular expression levels from microarray data of 6 Drosophila species centered by their medians. The modular expression divergence (inter-quartile region) becomes minimal during the fly phylotypic stage (brown, 8-10 hours). In the middle panel, the boxes show the modular expression levels of 16 conserved modules in flyusing only modENCODE RNA-seq data. The modular expression levels have the minimal variance across 16 conserved modules at the phylotypic stage (brown) in *D. melanogaster*. The bottom panel shows the modular expression correlations over a sliding 2-hour window (Pearson correlation per 5 stages, middle time of two-hour period in x-axis) among 16 modules in worm. We found that the modular correlations (median shown as bar height in y-axis) are highest during the worm phylotypic stages (brown), 6-8 hours. (More details on all parts of this in Figure S4.)

## Fig 7 - Worm & Fly Developmental Stage Alignment

(A) Alignment of worm and fly developmental stages based on all worm-fly orthologs. (B) Alignment of worm and fly developmental stages based on hourglass genes only. The alignment exhibits a gap, perfectly matching the phylotypic stage . Part (C) shows the key aligned stages from parts (A) and (B). Worm “early embryo” and “late embryo” stages are matched with fly “early embryo” and “late embryo” respectively in the “lower diagonal” set of matches, and they are also matched with fly “L1” and “prepupa-pupa” stages respectively in the “upper diagonal” set of matches.

## Fig 8 – TF and Histone Models for Gene Expression

TF and histone models for predicting gene expression. (A) Correlation around the TSS between gene expression and the signals for RNA Polymerase II binding or H3K4me3. (B) Genes are categorized into stalling (high binding, low expression), bursting (low binding, high expression) and normal (the rest). The grid shows the distribution for orthologous genes; the majority of them (75%) are normal across all three organisms. (C) Heat map showing the normalized correlation of binding with expression for various histone marks and transcription factors. For each of the three organisms, correlations are reported in 100 bp bins in a +/- 2kb window centered on the TSS for the same histone marks and two representative TFs. (D) The accuracy the TF model and histone-mark (HM) model for predicting expression of protein-coding genes and non-coding RNAs in each of 40 bins centered around the TSS. (E) The relative importance of histone marks in organism-specific HM models and in the universal HM model. The prediction accuracy of the TF model is also presented as a function as the number of independent TFs that are included. (F) Cross-organism prediction accuracy of the HM models and prediction accuracy of the universal HM model in human, worm, and fly.

$$\_{}^{}\_{}^{}\frac{\_{}^{}\_{}\_{}\_{}\_{}^{}}{\_{}^{}\_{}^{}}$$