Published in final edited form as:

Nature. 2014 August 28; 512(7515): 445–448. doi:10.1038/nature13424.

Comparative Analysis of the Transcriptome across Distant Species

A full list of authors and affiliations appears at the end of the article.

The transcriptome is the readout of the genome. Identifying common features in it across distant species can reveal fundamental principles. To this end, 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 metazoan phyla, extending beyond earlier within-phylum transcriptome comparisons and revealing ancient, conserved features[1,2,3,4,5,6]. Specifically, we discovered co-expression modules shared across animals, many of which are enriched in developmental genes. Moreover, we used expression patterns to align the stages in worm and fly development, finding a novel pairing between worm embryo and fly pupae, in addition to the expected embryo-to-embryo and larvae-to-larvae pairings. Furthermore, we found that the extent of non-canonical, non-coding transcription is similar in each organism, per base-pair. Finally, we found in all three organisms the gene-expression levels, both coding and non-coding, can be quantitatively predicted from chromatin features at the promoter using a "universal model," based on a single set of organism-independent parameters.

Our comparison used the ENCODE-modENCODE RNA resource (Fig. ED1). This resource comprises: (1) deeply sequenced RNA-Seq data from many distinct samples from all three organisms; (2) comprehensive annotation of transcribed elements and (3) uniformly 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. In total, the resource contains 575 different experiments containing >67B sequence reads. It encompasses many different RNA types, including poly(A)+, poly(A)- and ribosomal-RNA-depleted RNA and short and long RNA.

Correspondence and requests for materials should be addressed to the corresponding senior authors (cmptxn@gersteinlab.org).

*These authors contributed equally to this work.

#Co-senior authors.

Author Contributions

Work on the paper was divided between data production and analysis. The analysts were JR, KKY, DW, CC, JBB, CS, JJL, BP, AOH, MOD, SD, RPA, BHA, RKA, PJB, NPB, CD, AD, GF, AF, RG, JH, LH, HH, TH, RRK, JL, JL, ZL, AM, RM, PP, DP, AS, KW, KY, YZ and HZ (names are sorted according to order appearing in author list). The data producers were CAD, LH, KB, MEB, BWB, LC, PC, JD, BE, MF, GG, PG, AH, RAH, CH, CH, SJ, DK, MK, TCK, EL, EL, MM, GM, RM, GM, DMM, BO, SO, NP, VR, AR, GR, AS, GIS, FS, FJS, WCS, MHS, PS, KLW, JW, CX, LY and CZ. Significantly larger efforts are ascribed to the joint first authors. EAF, PJG, MJP: The role of the NIH Project Management Group was limited to coordination and scientific management of the modENCODE and ENCODE consortia. Overall project management was carried out by the senior authors MBG, RW, TRG, SEC, BRG and SEB.

The authors declare no competing financial interests.

The annotation in the resource represents capstones for the decade-long efforts in human, worm, and fly. The new annotation sets have numbers, sizes and families of protein-coding genes similar to previous compilations; however, the number of pseudogenes and annotated ncRNAs differ (Figs. ED2, ED3, S1). Also, the number of splicing events is greatly increased, resulting in a concomitant increase in protein complexity. We find the proportion of the different types of alternative splicing (e.g., exon skipping or intron retention) is generally similar across the three organisms; however, skipped exons predominate in human while retained introns are most common in worm and fly[7] (Figs. ED4, S1 and Table S1).

A fraction of the transcription comes from genomic regions not associated with standard annotations, representing “non-canonical transcription” (Table S2)[8]. Using a minimum-run/maximum-gap algorithm to process reads mapping outside of protein-coding transcripts, pseudogenes and annotated ncRNAs, we identified read clusters, i.e. transcriptionally active regions (TARs). Across all three genomes we found roughly one third of the bases gives rise to TARs, representing non-canonical transcription (Fig. ED3). To determine the extent that this transcription represents an expansion of the current established classes of ncRNAs, we identified the TARs most similar to known annotated ncRNAs using a supervised classifier[9] (Fig. S2, Table S2). We validated the classifier’s predictions using RT-PCR, demonstrating high accuracy. Overall, the predictions encompass 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 [8,10,11,12,13], finding statistically significant overlaps (Fig. ED5, Table S2).

Given the uniformly processed nature of the data and annotations, we were able to make comparisons across organisms. First, we built co-expression modules, extending earlier analysis[14](Fig. 1a). To detect modules consistently across the three species, we combined across-species orthology and within-species co-expression relationships. In the resulting multilayer network we searched for dense subgraphs (modules), using simulated annealing[15,16]. We found some modules dominated by a single species, whereas others contain genes from two or three. As expected, the modules with genes from multiple species are enriched in orthologs. Moreover, a phylogenetic analysis shows that the genes in such modules are more conserved across 56 diverse animal species (Figs. ED6, S3). 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. 1a, Table S3). Finally, we annotated many TARs based on correlating their expression profiles with these modules (Fig. ED5).

Next, we used the expression profiles of orthologous genes to align the developmental stages in worm and fly (Fig. 1b, ED7). 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 among 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. First, they match in the expected one-to-one fashion to the fly (i.e. embryos-to-embryos, larvae-to-larvae). However, worm late embryonic stages also match fly pupal stages, suggesting a shared expression

program between embryogenesis and metamorphosis. The ~50 stage-associated genes involved in this dual alignment are enriched in functions such as ion transport and cation-channel activity (Table S3).

To gain further insight into the stage alignment, we examined our conserved modules in terms of the "hourglass hypothesis", which posits that all animals 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[4,5,17]. For genes in 12 of the 16 modules, we observed canonical hourglass behavior, i.e. *inter*-organism expression divergence across closely related fly species during development is minimal[5](Fig. S3). Moreover, we find a subset of TARs also exhibit this "hourglass" behavior (Fig. S2). Beyond looking at *inter*-species divergence, we also investigated the *intra*-species divergence within just *D. melanogaster* and *C. elegans*. Strikingly, we observed that divergence of gene expression between modules is minimized during the worm and fly phylotypic stages (Fig. 1c). 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 before and after this. One can, in fact, directly see this coordination as a local maximum in between-module correlations for the worm (Fig. ED6). Finally, using genes from just the 12 "hourglass modules," we found that the alignment between worm and fly stages becomes stronger (Fig. 1b, S3). The alignment shows a gap where no changes are observed, perfectly matching the phylotypic stage.

The uniformly processed and matched nature of the transcriptome data also facilitates integration with upstream factor-binding and chromatin-modification signals. We investigated the degree to which these upstream signals can quantitatively predict gene expression and how consistent this prediction is across organisms. Similar to previous reports[11,18,19], we found consistent correlations, around the TSS, in each of the three species between various histone-modification signals and the expression level of the downstream gene: H3K4me1, H3K4me2, H3K4me3 and H3K27ac are positively correlated, whereas H3K27me3 is negatively correlated (Figs. 2, ED8, S4). Then for each organism, we integrated these individual correlations into a multivariate, statistical model, obtaining high accuracy in predicting expression for protein-coding genes and ncRNAs. The promoter-associated marks, H3K4me2 and H3K4me3, consistently have the highest contribution to the model.

A similar statistical analysis with TFs showed the correlation between gene expression and transcription-factor (TF) binding to be the greatest at the TSS, positively for activators and negatively for repressors (Fig. ED8). Integrated TF models in each organism also achieved high accuracy for protein-coding genes and ncRNAs, with as few as five TFs necessary for accurate predictions (Fig. ED9). This, perhaps, 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 ones, likely reflecting the fact that histone modifications are spread over broader regions, including the gene body, whereas most TFs bind near the promoter.

Finally, we constructed a "universal model," containing a single set of organism-independent parameters (Figs 2, S4). 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 that signals for most human enhancers are not near the TSS. Using the same set of organism-independent parameters derived from training on protein-coding genes, the universal model can also accurately predict ncRNA expression.

Overall, our comparison of the transcriptomes of three phylo-genetically distant metazoans highlights fundamental features of transcription conserved across animal phyla. First, there are ancient co-expression modules across organisms, many of which are enriched 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. The expression clustering also aligns developmental stages between worm and fly, revealing shared expression programs between embryogenesis and metamorphosis. Finally, we were able to build a single model that could predict transcription in all three organisms 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 divergent model organisms to human to highlight conserved biological principles (and dis-entangle them from lineage-specific adaptations).

Methods

Detailed methods are in the supplement. (See first section of this for a guide.) Data sets described here can be obtained from the ENCODE project website at http://www.encodeproject.org via accession number ENCSR145VDW. More detail on data availability is in section F of the supplement.

Extended Data

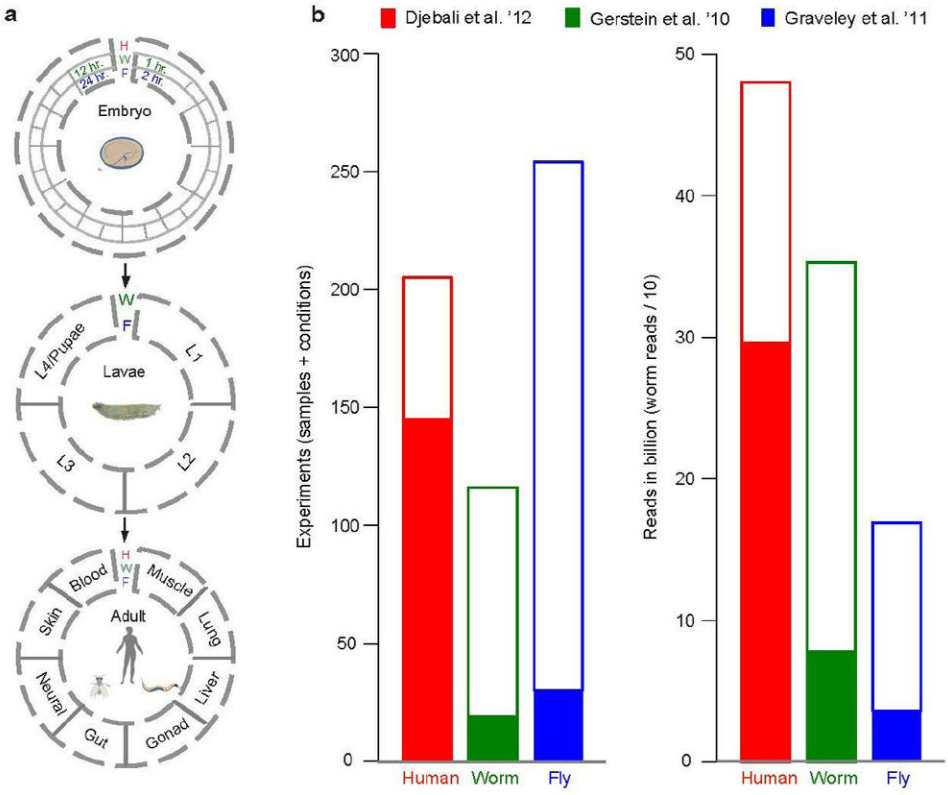

Fig ED1. Overview of the data

(A) Schematic of the RNA-seq data generated for human (red), worm (green), and fly (blue), showing how it samples developmental stages and various tissues and cell lines. (B) The number and size of data sets generated. The amount of new data beyond that in the previous ENCODE publications[8, 11, 20] is indicated by white bars, with previous ENCODE data indicated by solid bars. (See Supplement section B.2 for a detailed description of these data.)

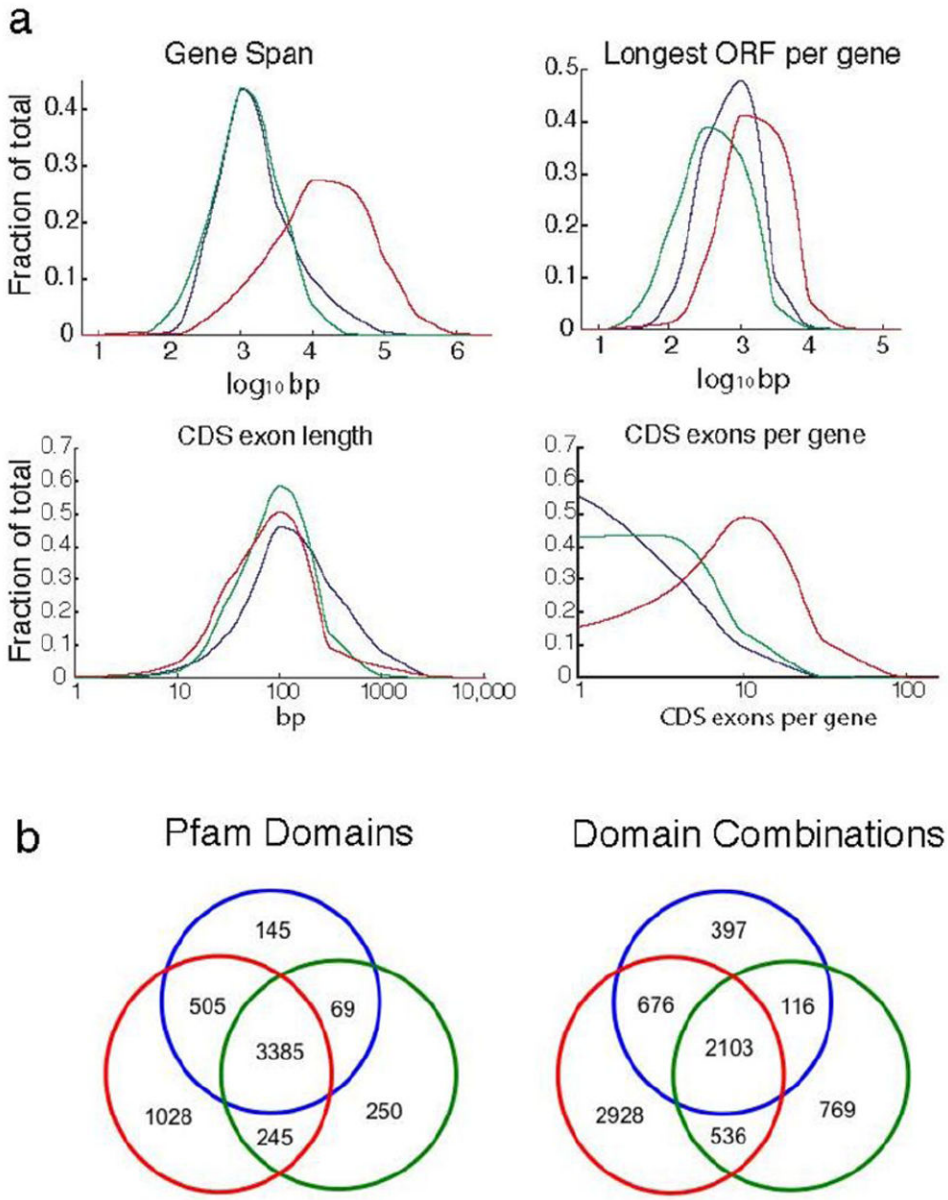

Fig ED2. Summary plots for the protein-coding gene annotations

(A) Distributions of key summary statistics - gene span, longest ORF per gene, CDS exon length, and CDS exons per gene; note that the x axes are in log scale. Both fly and worm genes span similar genomic lengths while human genes span larger regions (mostly due to the size of human introns). (B) Left: Venn diagram of protein domains (from the Pfam database version 26.0) present in annotated protein-coding genes in each species. Right: Shared domain combinations. (For more information on domain combinations, see Fig S1h and Supplement section B.4.1.)

				Elements	Human Genome Coverage Kb	%	Elements	Worm Genome Coverage Kb	%	Elements	Fly Genome Coverage Kb	%
Sequenced Genome	mRNAs (exons)			20,007	86,560	3.0	21,192	34,437	34.3	13,940	35,970	28.0
	Pseudogenes			11,216	27,089	0.95	881	1,343	1.3	145	155	0.12
	Annotated ncRNAs			22,154	17,77	0.62	41,466	2,611	2.6	2,155	3,279	2.6
		Comparable ncRNAs	miRNAs	1,756	162	0.006	221	20	0.02	236	22	0.02
			tRNAs	624	47	0.002	609	45	0.04	314	22	0.02
			snoRNAs	1,521	168	0.006	141	16	0.02	287	34	0.03
			snRNAs	1,944	210	0.007	114	14	0.01	47	7	0.006
			lncRNAs	10,840	10,581	0.37	233	184	0.18	852	868	0.68
	Regions Excluding mRNAs, Pseudogenes & Anno. ncRNAs			283,816	2,731,811	95.5	143,372	63,520	63.3	60,108	89,445	69.6
		Transcripton Detected (TARs)		708,253	916,401	32.0	232,150	37,029	36.9	83,618	44,256	34.5
			Supervised Predictions	104,016	13,835	0.48	2,525	392	0.39	599	164	0.13

Fig ED3. Summary of annotated ncRNAs, TARs, and ncRNA predictions in each species

The number of elements, the base pairs covered and the fraction of the genome for each class (see also Supplement section C). There are comparable numbers of tRNAs in humans and worms but about half as many in fly. While the number of lncRNAs in human is more than an order of magnitude greater than in either worms or flies, the fractional genomic coverage in all three species is similar. Finally, humans have at least 5-fold more miRNAs, snoRNAs and snRNAs compared to worm or fly. The fraction of the genome covered by TARs (highlighted squares) for each species is similar. A large amount of non-canonical transcription occurs in the introns of annotated genes, presumably representing a mixture of unprocessed mRNAs and internally initiated transcripts. The remaining non-canonical transcription (249Mb, 16Mb, and 14Mb in human, worm, and fly) is intergenic and occurs at low levels, comparable to that observed for introns (Table S2). Overall, the fraction of the genome transcribed -- including intronic, exonic, and non-canonical transcription -- is consistent with that previously reported for human despite the methodological differences in the analysis (Fig. S2, Supplement section C).

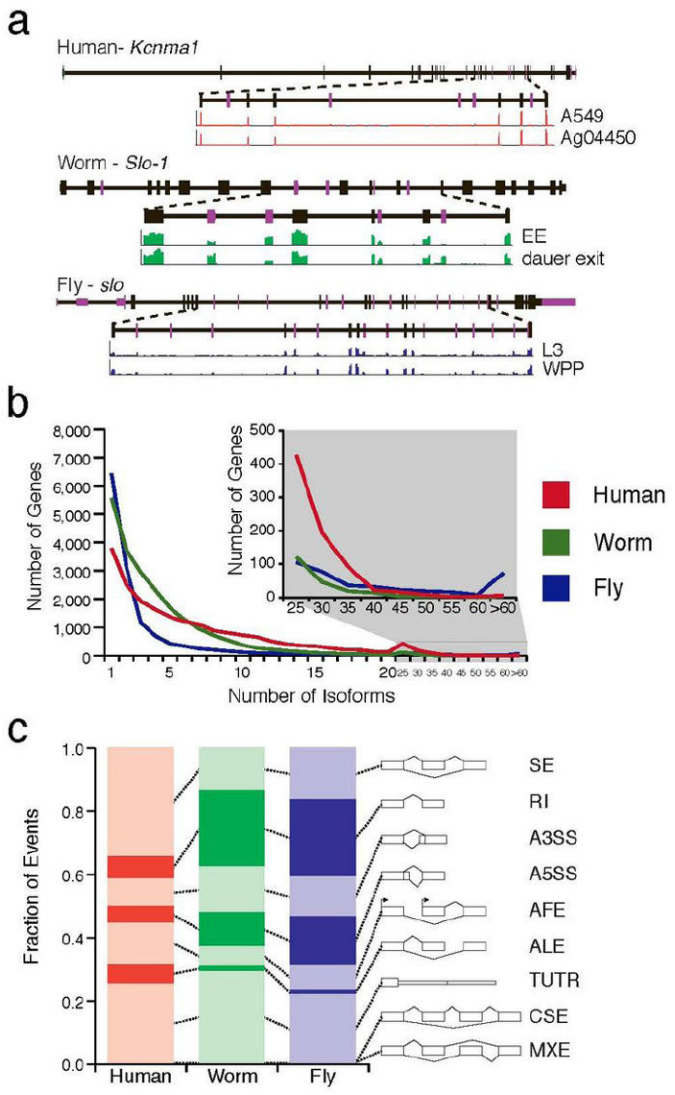

Fig ED4. Analysis of Alternative Splicing

(A) Representative orthologous genes do not share the same exon/intron structure or alternative splicing across species. (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”. (See Supplement section B.5 for a further discussion of splicing.)

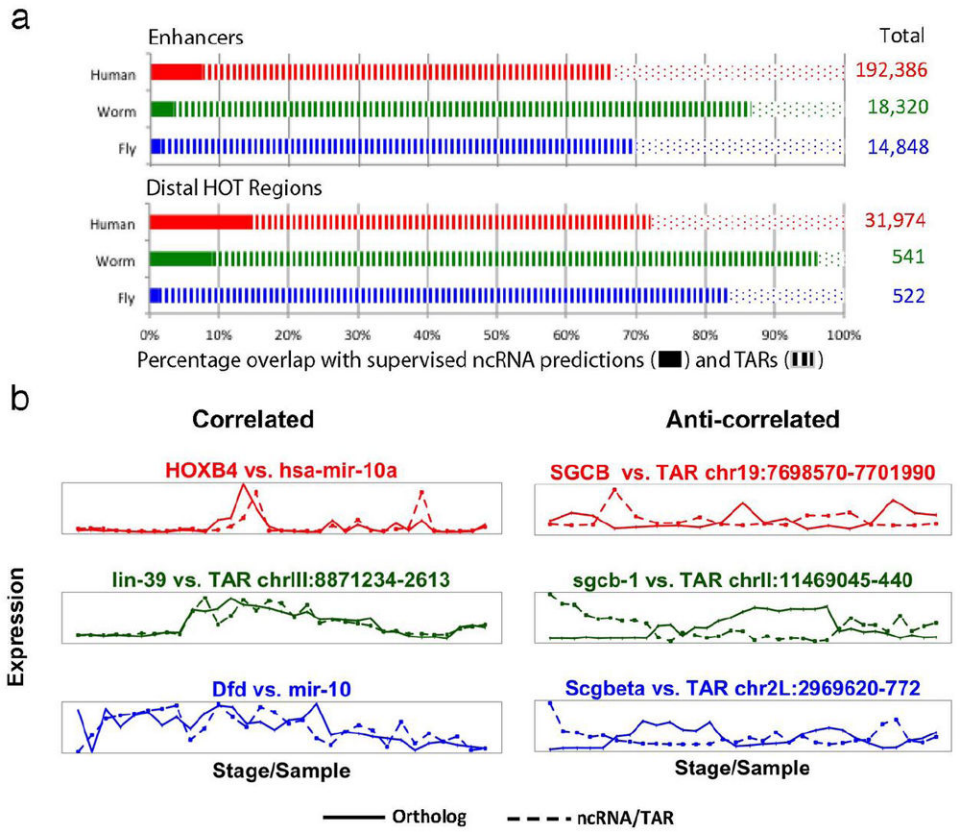

Fig ED5. Characterizing Non-canonical Transcription

(A) The overlap of enhancers and distal HOT regions with supervised ncRNA predictions and TARs in human, worm, and fly. The 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. (B) The left side highlights ncRNA/TARs that are highly correlated with corresponding HOX orthologues in human (HOXB4), worm (lin-39), and fly (Dfd). The expression of mir-10 correlates strongly with Dfd in fly ($r=0.66$, $p<6e-4$ in fly), as does mir-10a in human, which correlates strongly with HOXB4 ($r=0.88$, $p<2e-9$). A TAR (chrIII:8871234-2613) strongly correlates with *lin-39* ($r=0.91$, $p<4e-13$) in worm. The right side shows 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 ($r=-0.91$, $p<3e-16$), sgcb-1 ($r=-0.86$, $p<2e-7$), and Scgb ($r=-0.82$, $p<4e-8$), respectively. (More details on all parts of this figure are in Supplement section C and Table S2.)

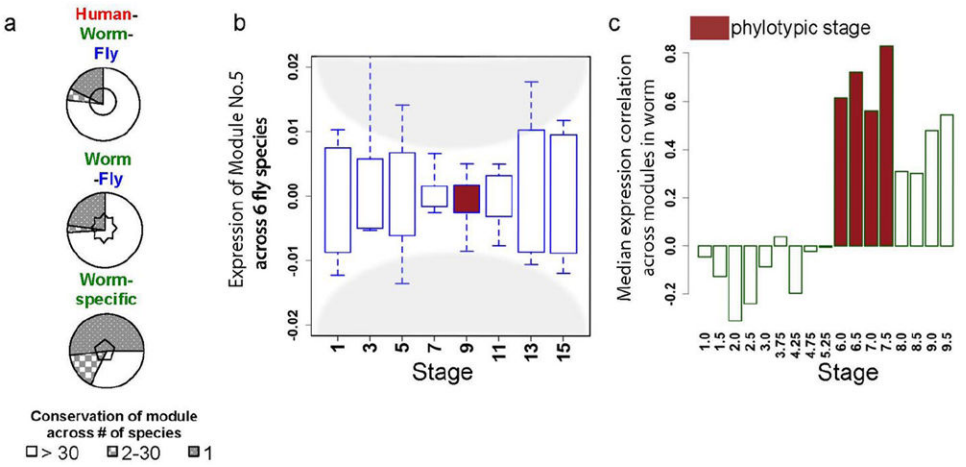

Fig ED6. Details on Expression Clustering

(A) Pie charts showing gene conservation across 56 Ensembl species for the blocks in the Fig. 1 heatmap enclosed with the same symbol (i.e. pentagon here matches pentagon in Fig. 1a). Overall, species-specific modules tend to have fewer orthologs across 56 Ensembl species. (B) The expression levels of a conserved module (Module No. 5) 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). (C) 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 are plotted. The modular correlations (median shown as bar height in y-axis) are highest during the worm phylotypic stages (brown), 6-8 hours. One can, in fact, directly see this coordination as a local maximum in the between-module correlation for the worm, which has a more densely sampled developmental time course. (This figure provides more detail on Fig. 1a and 1c. More details on all parts of this figure are in Supplement section D and Figure S3.)

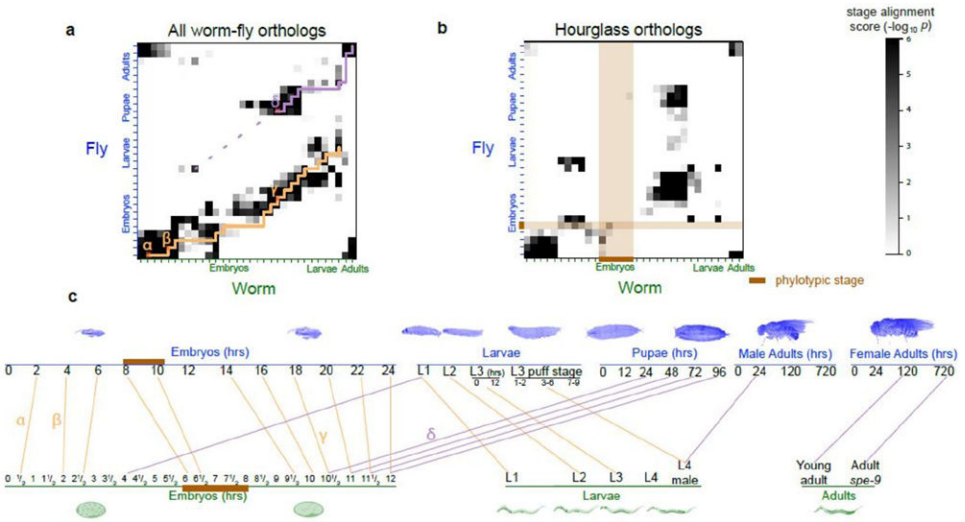

Fig ED7. Details on Stage Alignment

This figure provides further detail beyond Fig. 1b. (A) An alignment of worm and fly developmental stages based on all worm-fly orthologs (11,403 pairs, including one-to-one, one-to-many, many-to-many pairs). (B) Alignment of worm and fly developmental stages based on just worm-fly hourglass orthologs. Note the prominent gap in the aligned stages coincides with the worm and fly phylotypic stages (brown band). This make sense: since the expression values of genes in all hourglass modules converge at the phylotypic stage, no hourglass genes can be phylotypic-stage specific, and hence, the gap. (C) Key aligned stages from part (A). The correspondence between parts (A) and (C) is indicated by the small Greek letters. 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. (More details on all parts of this figure are in Supplement section D.4 and Table S3.)

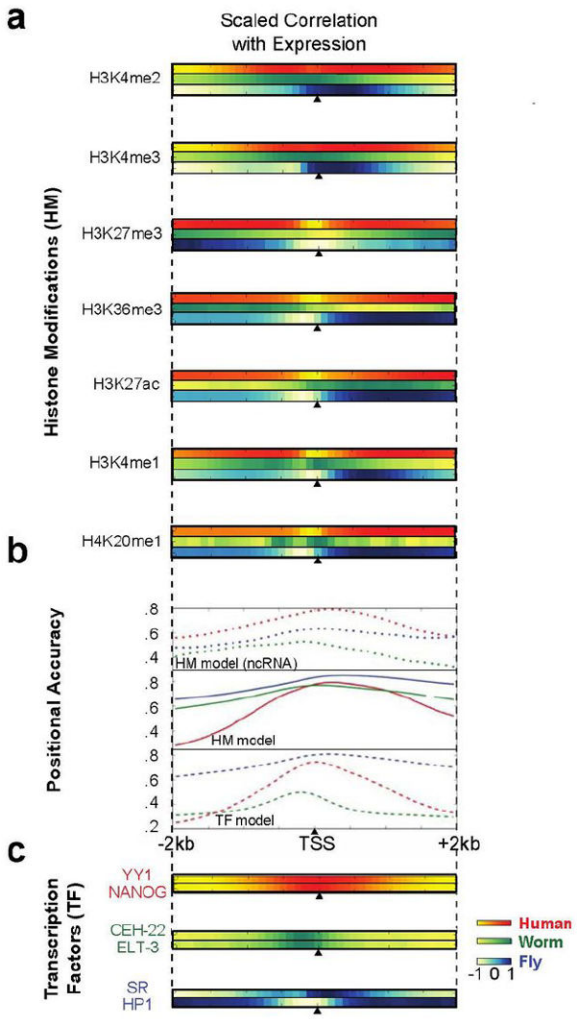

Fig ED8. Further Detail on Statistical Models for Predicting Gene Expression

This figure provides further information beyond that in Fig. 2. Binding/expression correlations of (A) various histone marks and (C) TFs. For instance, H3K36me3 shows positive correlation in worm and fly, but weak negative correlation in human at the promoter, with positive correlation over the gene body. (B) The positional accuracy from the TF and histone-mark models for predicting mRNA and ncRNA expression about the TSS. (More details on all parts of this figure are in Supplement section E and Fig. S4.)

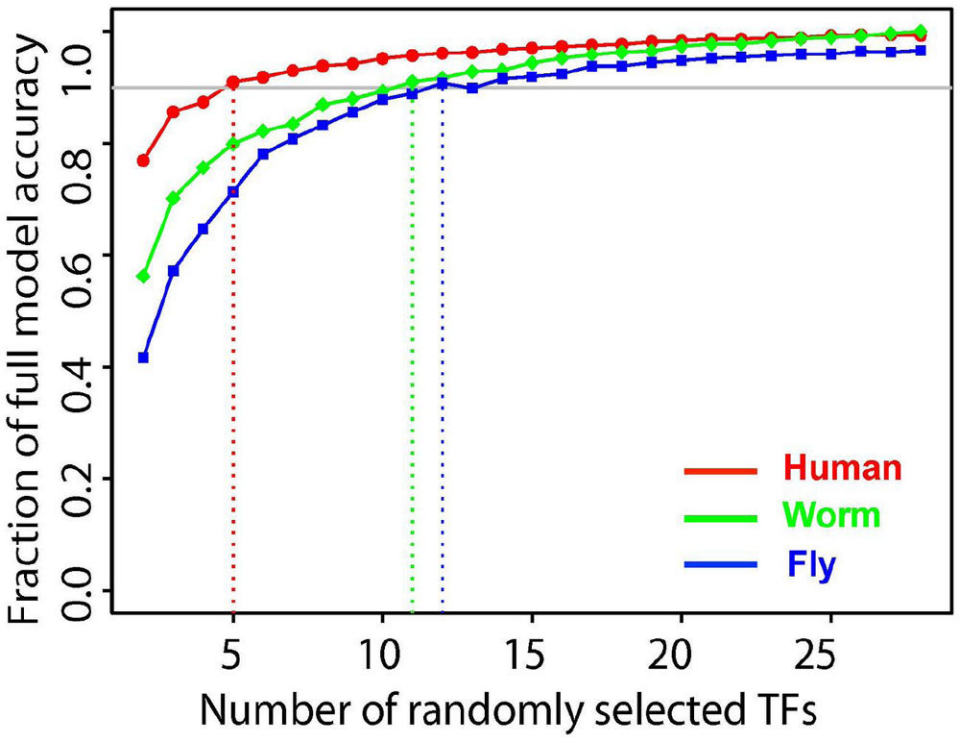

Fig ED9. Average predictive accuracy of models with different number of randomly selected TFs

We randomly selected n TFs as predictors and examined the predictive accuracy by cross-validation, where n varied from 2 to 28. The curve shows the average predictive accuracy (Fig. S4 indicates the standard deviation of all models with the same number of predictors). Surprisingly, models with as few as 5 TFs have predictive accuracy. This may reflect an intricate, correlated structure to regulation. However, it could also be that open chromatin is characteristic of gene expression and TFs bind somewhat indiscriminately. (More details on all parts of this figure are in Supplement section E.)

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Authors

Mark B. Gerstein[1,2,3,*,#], Joel Rozowsky[1,2,*], Koon-Kiu Yan[1,2,*], Daifeng Wang[1,2,*], Chao Cheng[4,5,*], James B. Brown[6,7,*], Carrie A Davis[8,*], LaDeana Hillier[9,*], Cristina Sisu[1,2,*], Jingyi Jessica Li[7,10,11,*], Baikang Pei[1,2,*], Arif O. Harmanci[1,2,*], Michael O. Duff[12,*], Sarah Djebali[13,14,*], Roger P. Alexander[1,2], Burak H. Alver[15], Raymond Auerbach[1,2], Kimberly Bell[8], Peter J. Bickel[7], Max E. Boeck[9], Nathan P. Boley[6,16], Benjamin W. Booth[6], Lucy Cherbas[17,18], Peter Cherbas[17,18], Chao Di[19], Alex Dobin[8], Jorg Drenkow[8], Brent Ewing[9], Gang Fang[1,2], Megan Fastuca[8], Elise A. Feingold[20], Adam Frankish[21], Guanjun Gao[19], Peter J. Good[20], Roderic Guigó[13,14], Ann Hammonds[6], Jen Harrow[21], Roger A. Hoskins[6], Cédric Howald[22,23], Long Hu[19], Haiyan Huang[7], Tim J. P. Hubbard[21,24], Chau Huynh[9], Sonali Jha[8], Dionna

Kasper[25], Masaomi Kato[26], Thomas C. Kaufman[17], Robert R. Kitchen[1,2], Erik Ladewig[27], Julien Lagarde[13,14], Eric Lai[27], Jing Leng[1,2], Zhi Lu[19], Michael MacCoss[9], Gemma May[12,28], Rebecca McWhirter[29], Gennifer Merrihew[9], David M. Miller[29], Ali Mortazavi[30,31], Rabi Murad[30,31], Brian Oliver[32], Sara Olson[12], Peter J. Park[15], Michael J. Pazin[20], Norbert Perrimon[33,34], Dmitri Pervouchine[13,14], Valerie Reinke[25], Alexandre Reymond[22], Garrett Robinson[7], Anastasia Samsonova[33,34], Gary I. Saunders[21,35], Felix Schlesinger[8], Anurag Sethi[1,2], Frank J. Slack[26], William C. Spencer[29], Marcus H. Stoiber, Pnina Strasbourger[9], Andrea Tanzer[36,37], Owen A. Thompson[9], Kenneth H. Wan[6], Guilin Wang[25], Huaien Wang[8], Kathie L. Watkins[29], Jiayu Wen[27], Kejia Wen[19], Chenghai Xue[8], Li Yang[12,38], Kevin Yip[39,40], Chris Zaleski[8], Yan Zhang[1,2], Henry Zheng[1,2], Steven E. Brenner[41,42,#], Brenton R. Graveley[12,#], Susan E. Celniker[6,#], Thomas R Gingeras[8,#], and Robert Waterston[9,#]

Affiliations

[1]Program in Computational Biology and Bioinformatics, Yale University, Bass 432, 266 Whitney Avenue, New Haven, Connecticut 06520, USA [2]Department of Molecular Biophysics and Biochemistry, Yale University, Bass 432, 266 Whitney Avenue, New Haven, Connecticut 06520, USA [3]Department of Computer Science, Yale University, 51 Prospect St, New Haven, Connecticut 06511, USA [4]Department of Genetics, Geisel School of Medicine at Dartmouth, Hanover, New Hampshire 03755, USA [5]Institute for Quantitative Biomedical Sciences, Norris Cotton Cancer Center, Geisel School of Medicine at Dartmouth, Lebanon, New Hampshire 03766, USA [6]Department of Genome Dynamics, Lawrence Berkeley National Laboratory, Berkeley, California 94720, USA [7]Department of Statistics, University of California, Berkeley, 367 Evans Hall, Berkeley, CA 94720-3860, USA [8]Functional Genomics, Cold Spring Harbor Laboratory, Cold Spring Harbor, N, Y, 11724 [9]Department of Genome Sciences and University of Washington School of Medicine, William H. Foege Bldg. S350D, 1705 N.E. Pacific Street, Box 355065 Seattle, Washington 98195-5065, USA [10]Department of Statistics, University of California, Los Angeles, CA 90095-1554, USA [11]Department of Human Genetics, University of California, Los Angeles, CA 90095-7088, USA [12]Department of Genetics and Developmental Biology, Institute for Systems Genomics, University of Connecticut Health Center, 400 Farmington Avenue, Farmington, CT 06030 USA [13]Centre for Genomic Regulation, Barcelona, Catalonia, Spain [14]Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain [15]Center for Biomedical Informatics, Harvard Medical School, 10 Shattuck St. Boston, Massachusetts 02115, USA [16]Department of Biostatistics, University of California, Berkeley, 367 Evans Hall, Berkeley, CA 94720-3860, USA [17]Department of Biology, Indiana University, 1001 E. 3rd Street, Bloomington, Indiana 47405-7005, USA [18]Center for Genomics and Bioinformatics, Indiana University, 1001 E. 3rd Street, Bloomington, Indiana 47405-7005, USA [19]MOE Key Lab of Bioinformatics, School of Life Sciences, Tsinghua University, Beijing, China 100084 [20]National Human Genome Research Institute, National Institutes of Health, 5635 Fishers Lane, Bethesda, Maryland 20892-9307, USA [21]Wellcome Trust Sanger Institute,

Wellcome Trust Genome Campus, Hinxton, UK [22]Center for Integrative Genomics, University of Lausanne, 1015 Lausanne, Switzerland [23]Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland [24]Medical and Molecular Genetics, King's College London, London WC2R 2LS, UK [25]Department of Genetics, Yale University School of Medicine, New Haven, Connecticut 06520-8005, USA [26]Department of Molecular, Cellular and Developmental Biology, PO Box 208103, Yale University, New Haven, Connecticut 06520, USA [27]Sloan-Kettering Institute, 1275 York Avenue, Box 252, New York, New York 10065, USA [28]Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA 15213 USA [29]Department of Cell and Developmental Biology, Vanderbilt University, 465 21st Avenue South, Nashville, Tennessee 37232-8240, USA [30]Developmental and Cell Biology, University of California, Irvine, CA 92697 [31]Center for Complex Biological Systems, University of California, Irvine, CA 92697 [32]Section of Developmental Genomics, Laboratory of Cellular and Developmental Biology, National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health, Bethesda MD 20892, USA [33]Department of Genetics and Drosophila RNAi Screening Center, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, USA [34]Howard Hughes Medical Institute, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, USA [35]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SD, UK [36]Bioinformatics and Genomics Programme, Center for Genomic Regulation, Universitat Pompeu Fabra (CRG-UPF), Barcelona, Catalonia, Spain [37]Institute for Theoretical Chemistry, Theoretical Biochemistry Group (TBI), University of Vienna [38]Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China [39]Hong Kong Bioinformatics Centre, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong [40]5 CUHK-BGI Innovation Institute of Trans-omics, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong [41]Department of Molecular and Cell Biology, University of California, Berkeley, California 94720, USA [42]Department of Plant & Microbial Biology, University of California, Berkeley, California 94720, USA

Acknowledgments

The authors thank the NHGRI and the mod/ENCODE projects for support. In particular, this work was funded by a contract from the National Human Genome Research Institute modENCODE Project, contract U01 HG004271 and U54 HG006944, to S.E.C. (Principal Investigator) and P.C., T.R.G., R.A.H., and B.R.G. (co-Principal Investigators) with additional support from R01 GM076655 (S.E.C.) both under Department of Energy contract no. DE-AC02-05CH11231, and U54 HG007005 to B.R.G. J.B.B.'s work was supported by NHGRI K99 HG006698 and DOE DE-AC02-05CH11231. Work in P.J.B.'s group was supported by the modENCODE DAC sub award 5710003102, 1U01HG007031-01 and the ENCODE DAC 5U01HG004695-04. Work in M.B.G.'s group was supported by NIH grants HG007000 and HG007355. Work in Bloomington was supported in part by the Indiana METACyt Initiative of Indiana University, funded by an award from the Lilly Endowment, Inc. Work in E.C.L.'s group was supported by U01-HG004261 and RC2-HG005639. P.J.P. acknowledges support from the National Institutes of Health (grant no. U01HG004258). We thank the HAVANA team for providing annotation of the human reference genome, whose work is supported by National Institutes of Health (grant no. 5U54HG004555), the Wellcome Trust (grant no. WT098051). R.G. acknowledges support from the Spanish Ministry of Education (grant BIO2011-26205). We also acknowledge use of the Yale University Biomedical High Performance Computing Center.

References

1. Brawand D, et al. The evolution of gene expression levels in mammalian organs. Nature. 2011; 478:343–8. [PubMed: 22012392]
2. Merkin J, Russell C, Chen P, Burge CB. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. Science. 2012; 338:1593–9. [PubMed: 23258891]
3. Barbosa-Morais NL, et al. The evolutionary landscape of alternative splicing in vertebrate species. Science. 2012; 338:1587–93. [PubMed: 23258890]
4. Levin M, Hashimshony T, Wagner F, Yanai I. Developmental milestones punctuate gene expression in the Caenorhabditis embryo. Dev Cell. 2012; 22:1101–8. [PubMed: 22560298]
5. Kalinka AT, et al. Gene expression divergence recapitulates the developmental hourglass model. Nature. 2010; 468:811–4. [PubMed: 21150996]
6. Simola DF, Francis C, Sniegowski PD, Kim J. Heterochronic evolution reveals modular timing changes in budding yeast transcriptomes. Genome Biol. 2010; 11:R105. [PubMed: 20969771]
7. Talerico M, Berget SM. Intron definition in splicing of small Drosophila introns. Mol Cell Biol. 1994; 14:3434–45. [PubMed: 8164690]
8. Djebali S, et al. Landscape of transcription in human cells. Nature. 2012; 489:101–8. [PubMed: 22955620]
9. Lu ZJ, et al. Prediction and characterization of noncoding RNAs in C. elegans by integrating conservation, secondary structure, and high-throughput sequencing and array data. Genome Res. 2011; 21:276–85. [PubMed: 21177971]
10. Boyle AP, et al. Comparative analysis of regulatory information and circuits across distant species. Nature. submitted.
11. Gerstein MB. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science. 2010; 330:1775–1787. [PubMed: 21177976]
12. modENCODE Consortium. Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science. 2010; 330:1787–97. [PubMed: 21177974]
13. Ho JWK, et al. modENCODE and ENCODE resources for analysis of metazoan chromatin organization. Nature. submitted.
14. Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. Science. 2003; 302:249–55. [PubMed: 12934013]
15. Kirkpatrick S, Gelatt CD Jr, Vecchi MP. Optimization by simulated annealing. Science. 1983; 220:671–80. [PubMed: 17813860]
16. Reichardt J, Bornholdt S. Detecting fuzzy community structures in complex networks with a Potts model. Phys Rev Lett. 2004; 93:218701. [PubMed: 15601068]
17. Domazet-Lošo T, Tautz D. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. Nature. 2010; 468:815–8. [PubMed: 21150997]
18. Karlić R, Chung H-R, Lasserre J, Vlahovicek K, Vingron M. Histone modification levels are predictive for gene expression. Proc Natl Acad Sci U S A. 2010; 107:2926–31. [PubMed: 20133639]
19. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012; 489:57–74. [PubMed: 22955616]
20. Graveley BR, et al. The developmental transcriptome of Drosophila melanogaster. Nature. 2011; 471:473–9. [PubMed: 21179090]

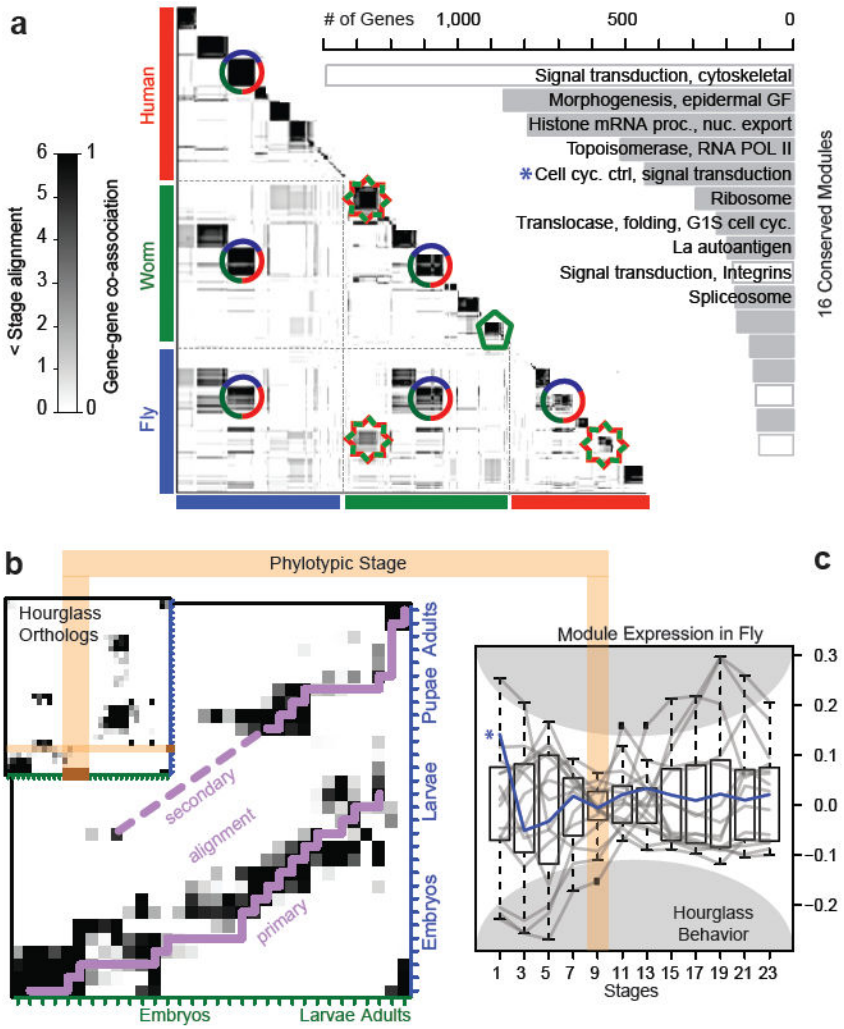

Fig 1. Expression Clustering

(A) Left: Human, worm, and fly gene-gene co-association matrix; darker coloring reflects the increased likelihood that a pair of genes are assigned to the same module. A dark block along the diagonal represents a group of genes within a species. If this is associated with an off-diagonal block then it is a cross-species module (e.g. a three-species conserved module is shown with a circle and a worm-fly module, with a star). However, if a diagonal block has no off-diagonal associations, then it forms a species-specific module (e.g. green pentagon). Right: The GO functional enrichment of genes within the 16 conserved modules 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 orthologs is more significant and exhibits a gap (brown) matching the phylotypic stage. (C) Normalized expression of the conserved modules in fly shows the smallest intra-organism divergence during the phylotypic stage (brown). (See Figs. ED 6 and 7 for further details.)

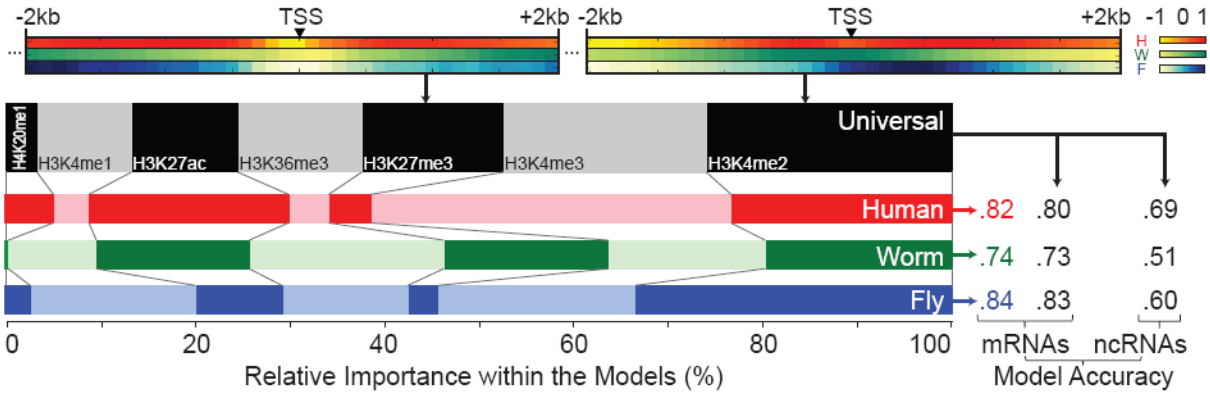

Fig 2. Histone Models for Gene Expression

Top: Normalized correlations of two representative histone marks with expression. Left: Relative importance of the histone marks in organism-specific models and the universal model. Right: Prediction accuracies (Pearson correlations all significant, p<1e-100) of the organism-specific and universal models. (See Figs. ED 8 and 9 for further details.)