LETTER

OPEN

doi:10.1038/nature13424

Comparative analysis of the transcriptome across distant species

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[6,16], 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]§ & Robert Waterston[9]§

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[1–6]. Specifically, we discover co-expression modules shared across animals, many of which are enriched in developmental genes. Moreover, we use expression patterns to align the stages in worm and fly development and find a novel pairing between worm embryo and fly pupae, in addition to the embryo-to-embryo and larvae-to-larvae pairings. Furthermore, we find that the extent of non-canonical, non-coding transcription is similar in each organism, per base pair. Finally, we find in all three organisms that 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 (Extended Data Fig. 1). This resource comprises: deeply sequenced RNA-sequencing (RNA-seq) data from many distinct samples from all three organisms; comprehensive annotation of transcribed elements; and uniformly processed, standardized analysis files, focusing on non-coding transcription and expression patterns. Where practical, these data sets match comparable samples across organisms and to other types of functional genomics data. In total, the resource contains 575 different experiments containing >67 billion sequence reads. It encompasses many different RNA types, including poly(A)+, poly(A)-, ribosomal-RNA-depleted, short and long RNA.

The annotation in the resource represents a capstone 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

[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 Street, 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, California 94720-3860, USA. [8]Functional Genomics, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA. [9]Department of Genome Sciences and University of Washington School of Medicine, William H. Foege Building S350D, 1705 Northeast Pacific Street, Box 355065 Seattle, Washington 98195-5065, USA. [10]Department of Statistics, University of California, Los Angeles, California 90095-1554, USA. [11]Department of Human Genetics, University of California, Los Angeles, California 90095-7088, USA. [12]Department of Genetics and Developmental Biology, Institute for Systems Genomics, University of Connecticut Health Center, 400 Farmington Avenue, Farmington, Connecticut 06030, USA. [13]Centre for Genomic Regulation, Doctor Aiguader 88, 08003 Barcelona, Catalonia, Spain. [14]Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, 08003 Barcelona, Catalonia, Spain. [15]Center for Biomedical Informatics, Harvard Medical School, 10 Shattuck Street, Boston, Massachusetts 02115, USA. [16]Department of Biostatistics, University of California, Berkeley, 367 Evans Hall, Berkeley, California 94720-3860, USA. [17]Department of Biology, Indiana University, 1001 East 3rd Street, Bloomington, Indiana 47405-7005, USA. [18]Center for Genomics and Bioinformatics, Indiana University, 1001 East 3rd Street, Bloomington, Indiana 47405-7005, USA. [19]MOE Key Lab of Bioinformatics, School of Life Sciences, Tsinghua University, Beijing 100084, China. [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, Cambridge CB10 1SA, UK. [22]Center for Integrative Genomics, University of Lausanne, Genopode building, Lausanne 1015, Switzerland. [23]Swiss Institute of Bioinformatics, Genopode building, Lausanne 1015, 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, Pennsylvania 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, California 92697, USA. [31]Center for Complex Biological Systems, University of California, Irvine, California 92697, USA. [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, Maryland 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), 08003 Barcelona, Catalonia, Spain. [37]Institute for Theoretical Chemistry, Theoretical Biochemistry Group (TBI), University of Vienna, Währingerstrasse 17/3/303, A-1090 Vienna, Austria. [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 and Microbial Biology, University of California, Berkeley, California 94720, USA.

*These authors contributed equally to this work.

§These authors jointly supervised this work.

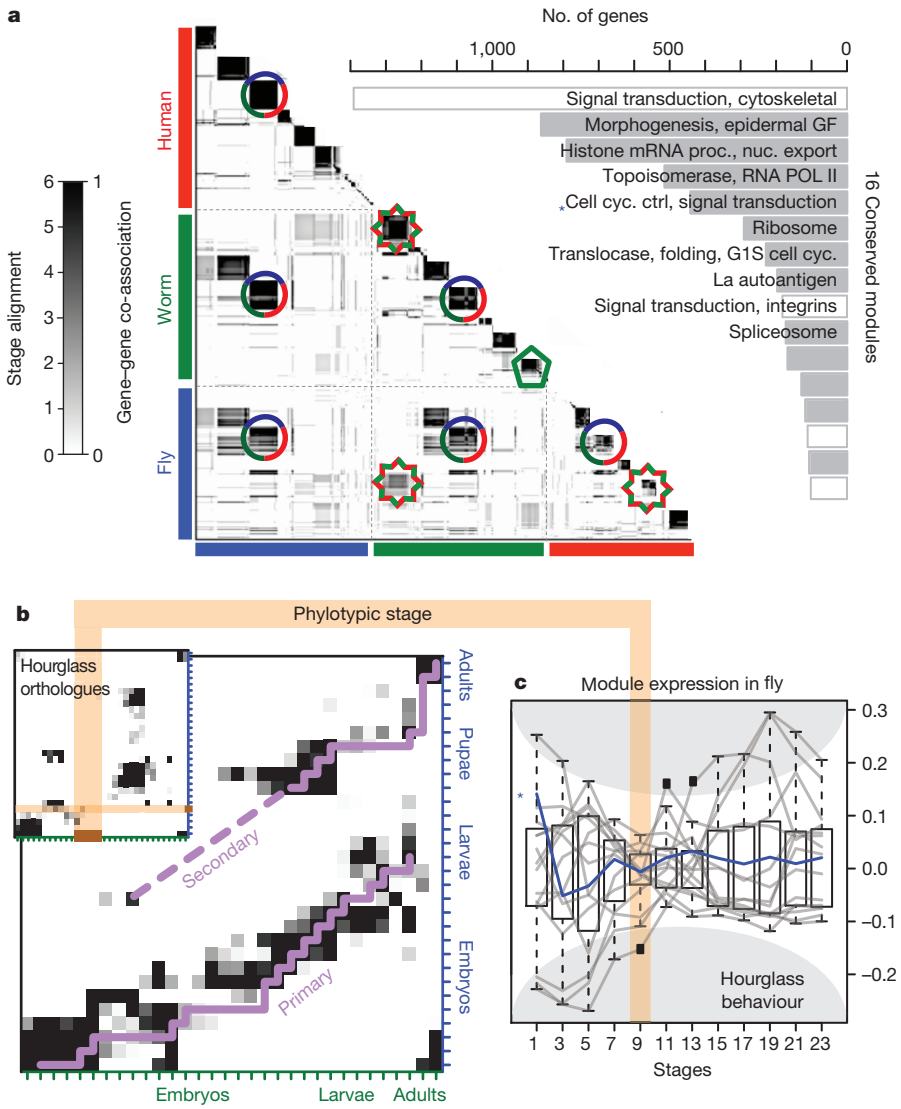


Figure 1 | Expression clustering. **a**, Left, human, worm and fly gene–gene co-association matrix; darker colouring reflects the increased likelihood that a pair of genes are assigned to the same module. A dark block along the diagonal represents a group of genes within a species. If this is associated with an off-diagonal block then it is a cross-species module (for example, 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 (for example, green pentagon). Right, the Gene Ontology functional enrichment of genes within the 16 conserved modules is shown. GF, growth factor; nuc., nuclear; proc., processing. **b**, Primary and secondary alignments of worm-and-fly developmental stages based on all worm–fly orthologues. Inset shows worm–fly stage alignment using only hourglass orthologues is more significant and exhibits a gap (brown) matching the phylotypic stage. The scale for the heat map in **b** is indicated on the left side of the scale in **a** (labelled stage alignment). **c**, Normalized expression of the conserved modules in fly shows the smallest intra-organism divergence during the phylotypic stage (brown). A representative module is indicated with a blue asterisk in **a** and **c**. (For further details see Extended Data Figs 5 and 6; ref. 20, related to the left part of **a**; and ref. 21, related to the bottom part of **b**.)

compilations; however, the number of pseudogenes and annotated non-coding RNAs differ (Extended Data Fig. 2, Extended Data Table 1 and Supplementary Fig. 1). 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 (for example, 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] (Extended Data Fig. 3, Supplementary Fig. 1 and Supplementary Table 1).

A fraction of the transcription comes from genomic regions not associated with standard annotations, representing 'non-canonical transcription' (Supplementary Table 2)[8]. Using a minimum-run–maximum-gap algorithm to process reads mapping outside of protein-coding transcripts, pseudogenes and annotated non-coding RNAs, we identified read clusters; that is, transcriptionally active regions (TARs). Across all three genomes we found roughly one-third of the bases gives rise to TARs or non-canonical transcription (Extended Data Table 1). To determine the extent that this transcription represents an expansion of the current established classes of non-coding RNAs, we identified the TARs most similar to known annotated non-coding RNAs using a supervised classifier[9] (Supplementary Fig. 2 and Supplementary Table 2). We validated the classifier's predictions using RT–PCR (PCR with reverse transcription), demonstrating high accuracy. Overall, these predictions encompass only a small fraction of all TARs, suggesting that most TARs have features distinct from annotated non-coding RNAs and that the majority of non-coding RNAs of established classes have already been identified. To shed further light on the possible roles of TARs we intersected them with enhancers and HOT (high-occupancy target) regions[8,10–13], finding statistically significant overlaps (Extended Data Fig. 4 and Supplementary Table 2).

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 orthologues. Moreover, a phylogenetic analysis shows that the genes in such modules are more conserved across 56 diverse animal species (Extended Data Fig. 5 and Supplementary Fig. 3). To focus on the cross-species conserved functions, we restricted the clustering to orthologues, arriving at 16 conserved modules, which are enriched in a variety of functions, ranging from morphogenesis to chromatin remodelling (Fig. 1a and Supplementary Table 3). Finally, we annotated many TARs based on correlating their expression profiles with these modules (Extended Data Fig. 4).

Next, we used expression profiles of orthologous genes to align the developmental stages in worm and fly (Fig. 1b and Extended Data Fig. 6). For every developmental stage, we identified stage-associated genes; that

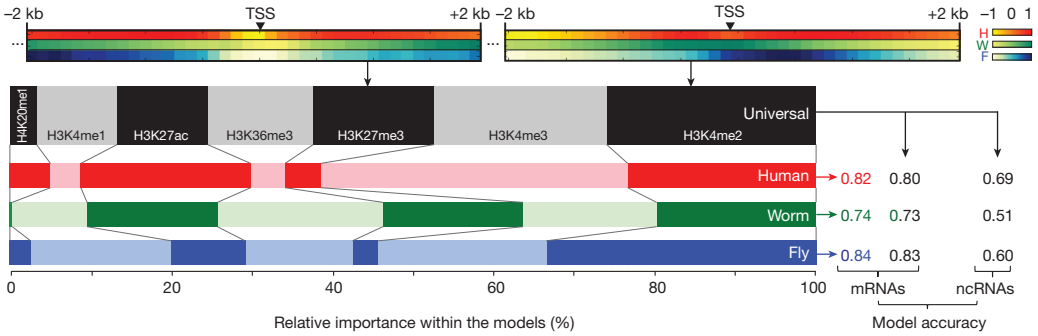


Figure 2 | Histone models for gene expression. Top, normalized correlations of two representative histone marks with expression. Left, relative importance of the histone marks in organism-specific models and the universal model. Right, prediction accuracies (Pearson correlations all significant, $P < 1 \times 10^{-100}$) of the organism-specific and universal models. (See Extended Data Figs 7 and 8 for further details.)

is, genes highly expressed at that 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. Notably, worm stages map to two sets of fly stages. First, they match in a co-linear fashion to the fly (that is, 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 approximately 50 stage-associated genes involved in this dual alignment are enriched in functions such as ion transport and cation-channel activity (Supplementary Table 3).

To gain further insight into the stage alignment, we examined our 16 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 behaviour; that is, inter-organism expression divergence across closely related fly species during development is minimal[5] (Supplementary Fig. 3). Moreover, we find a subset of TARs also exhibit this hourglass behaviour (Supplementary Fig. 2). Beyond looking at inter-species divergence, we also investigated the *intra*-species divergence within just *Drosophila melanogaster* and *Caenorhabditis elegans*. Notably, 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 expression signature before and after this. In fact it is possible to see this coordination directly as a local maximum in between-module correlations for the worm (Extended Data Fig. 5). Finally, using genes from just the 12 'hourglass modules', we found that the alignment between worm and fly stages becomes stronger (Fig. 1b and Supplementary Fig. 3); in particular it 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 transcription start site (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 (Fig. 2, Extended Data Fig. 7 and Supplementary Fig. 4). 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 non-coding RNAs. The promoter-associated marks, H3K4me2 and H3K4me3, consistently have the highest contribution to the model.

A similar statistical analysis with transcription factors showed the correlation between gene expression and transcription-factor binding to be the greatest at the TSS, positively for activators and negatively for repressors (Extended Data Fig. 7). Integrated transcription-factor models in each organism also achieved high accuracy for protein-coding genes and non-coding RNAs, with as few as five transcription factors necessary for accurate predictions (Extended Data Fig. 8). This perhaps reflects an intricate, correlated structure to regulation. The relative importance of the upstream regions is more peaked for the transcription-factor models than for the histone ones, likely reflecting the fact that histone modifications are spread over broader regions, including the gene body, whereas most transcription factors bind near the promoter.

Finally, we constructed a 'universal model', containing a single set of organism-independent parameters (Fig. 2 and Supplementary Fig. 4). 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 non-coding RNA expression.

Overall, our comparison of the transcriptomes of three phylogenetically 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 non-coding RNAs. Overall, our results underscore the importance of comparing divergent model organisms to human to highlight conserved biological principles (and disentangle them from lineage-specific adaptations).

METHODS SUMMARY

Detailed methods are given in the Supplementary Information. (See the first section of the Supplementary Information for a guide.) More details on data availability are given in section F of the Supplementary Information.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 10 April 2013; accepted 30 April 2014.

1. Brawand, D. *et al.* The evolution of gene expression levels in mammalian organs. *Nature* **478,** 343–348 (2011).
2. Merkin, J., Russell, C., Chen, P. & Burge, C. B. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. *Science* **338,** 1593–1599 (2012).
3. Barbosa-Morais, N. L. *et al.* The evolutionary landscape of alternative splicing in vertebrate species. *Science* **338,** 1587–1593 (2012).
4. Levin, M., Hashimshony, T., Wagner, F. & Yanai, I. Developmental milestones punctuate gene expression in the *Caenorhabditis* embryo. *Dev. Cell* **22,** 1101–1108 (2012).
5. Kalinka, A. T. *et al.* Gene expression divergence recapitulates the developmental hourglass model. *Nature* **468,** 811–814 (2010).

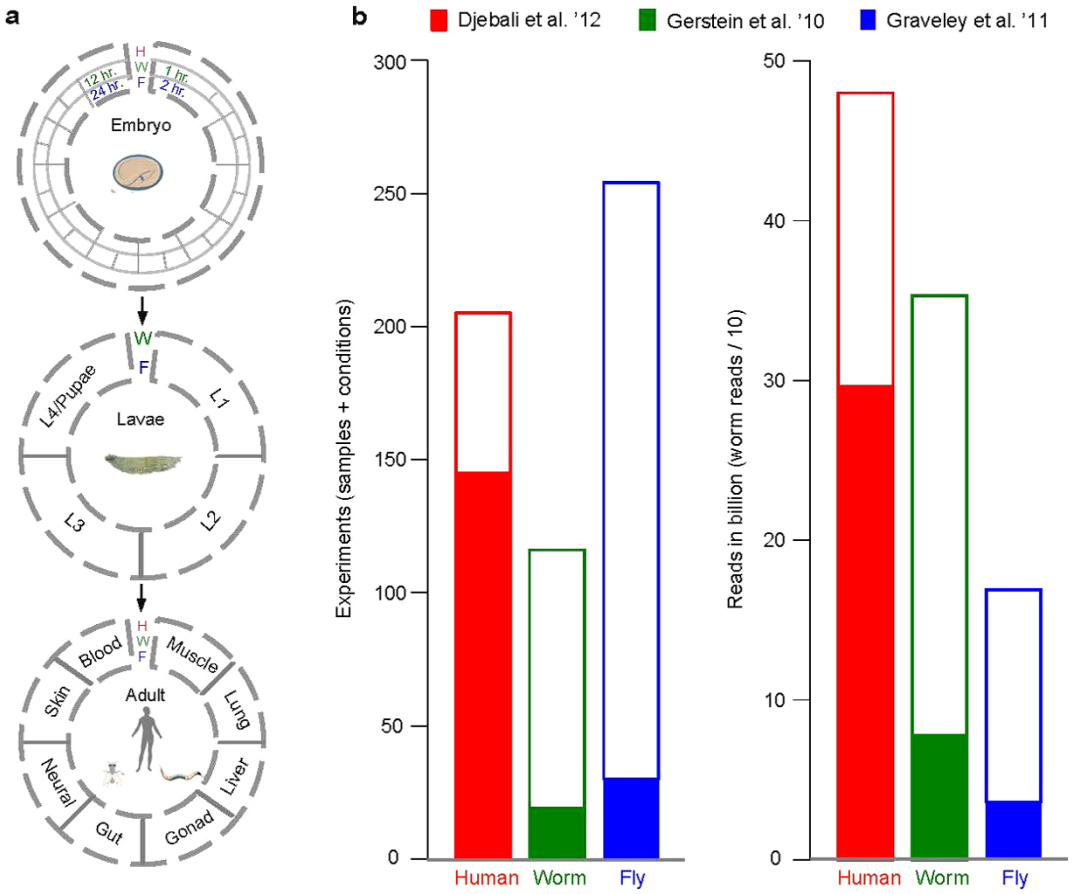
6. Simola, D. F., Francis, C., Sniegowski, P. D. & Kim, J. Heterochronic evolution reveals modular timing changes in budding yeast transcriptomes. *Genome Biol.* **11,** R105 (2010).
7. Talerico, M. & Berget, S. M. Intron definition in splicing of small *Drosophila* introns. *Mol. Cell. Biol.* **14,** 3434–3445 (1994).
8. Djebali, S. *et al.* Landscape of transcription in human cells. *Nature* **489,** 101–108 (2012).
9. Lu, Z. J. *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.* **21,** 276–285 (2011).
10. Boyle, A. P. *et al.* Comparative analysis of regulatory information and circuits across distant species. *Nature* http://dx.doi.org/10.1038/nature13668 (this issue).
11. Gerstein, M. B. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science* **330,** 1775–1787 (2010).
12. modENCODE Consortium, *et al.* Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* **330,** 1798–1797 (2010).
13. Ho, J. W. K. *et al.* modENCODE and ENCODE resources for analysis of metazoan chromatin organization. *Nature* http://dx.doi.org/10.1038/nature13497 (this issue).
14. Stuart, J. M., Segal, E., Koller, D. & Kim, S. K. A gene-coexpression network for global discovery of conserved genetic modules. *Science* **302,** 249–255 (2003).
15. Kirkpatrick, S., Gelatt, C. D., Jr & Vecchi, M. P. Optimization by simulated annealing. *Science* **220,** 671–680 (1983).
16. Reichardt, J. & Bornholdt, S. Detecting fuzzy community structures in complex networks with a Potts model. *Phys. Rev. Lett.* **93,** 218701 (2004).
17. Domazet-Lošo, T. & Tautz, D. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* **468,** 815–818 (2010).
18. Karlić, R., Chung, H.-R., Lasserre, J., Vlahovicek, K. & Vingron, M. Histone modification levels are predictive for gene expression. *Proc. Natl Acad. Sci. USA* **107,** 2926–2931 (2010).
19. ENCODE Project Consortium An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
20. Yan, K. K., Wang, D., Rozowsky, J., Zheng, H., Cheng, C. & Gerstein, M. OrthoClust: an orthology-based network framework for clustering data across multiple species. *Genome Biol.* **15,** R100 (2014).
21. Li, J. J., Huang, H., Bickel, P. J. & Brenner, S. E. Comparison of *D. melanogaster* and *C. elegans* developmental stages, tissues, and cells by modENCODE RNA-seq data. *Genome Res.* **24,** 1086–1101 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements The authors thank the NHGRI and the ENCODE and modENCODE 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. R.W.'s lab was supported by grant no. U01 HG 004263.

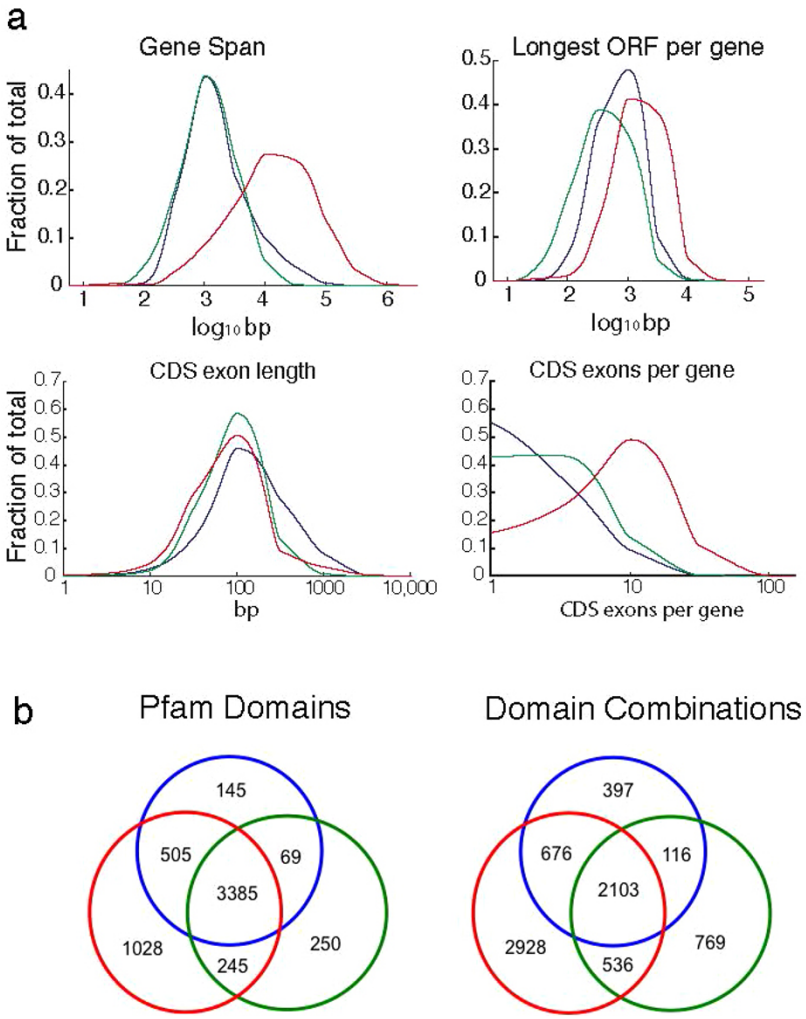
Author Contributions Work on the paper was divided between data production and analysis. The analysts were J.R., K.K.Y., D.W., C.C., J.B.B., C.S., J.J.L., B.P., A.O.H., M.O.D., S.D., R.P.A., B.H.A., R.K.A., P.J.B., N.P.B., C.D., A.D., G.F., A.F., R.G., J.H., L.H., H.H., T.H., R.R.K., J.L., J.L., Z.L., A.M., R.M., P.P., D.P., A.S., K.W., K.Y., Y.Z. and H.Z. (names are sorted according to their order in the author list). The data producers were C.A.D., L.H., K.B., M.E.B., B.W.B., L.C., P.C., J.D., B.E., M.F., G.G., P.G., A.H., R.A.H., C.H., C.H., S.J., D.K., M.K., T.C.K., E.L., E.L., M.M., G.M., R.M., G.M., D.M.M., B.O., S.O., N.P., V.R., A.R., G.R., A.S., G.I.S., F.S., F.J.S., W.C.S., M.H.S., P.S., K.L.W., J.W., C.X., L.Y. and C.Z. Substantially larger contributions were made by the joint first authors. The role of the NIH Project Management Group, E.A.F., P.J.G., M.J.P., was limited to coordination and scientific management of the modENCODE and ENCODE consortia. Overall project management was carried out by the senior authors M.B.G., R.W., T.R.G., S.E.C., B.R.G. and S.E.B.

Author Information Data sets described here can be obtained from the ENCODE project website at http://www.encodeproject.org/comparative via accession number ENCSR145VDW (alternate URL http://cmptxn.gersteinlab.org). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.B.G., R.W., T.R.G., S.E.C., B.R.G. or S.E.B. (cmptxn@gersteinlab.org).

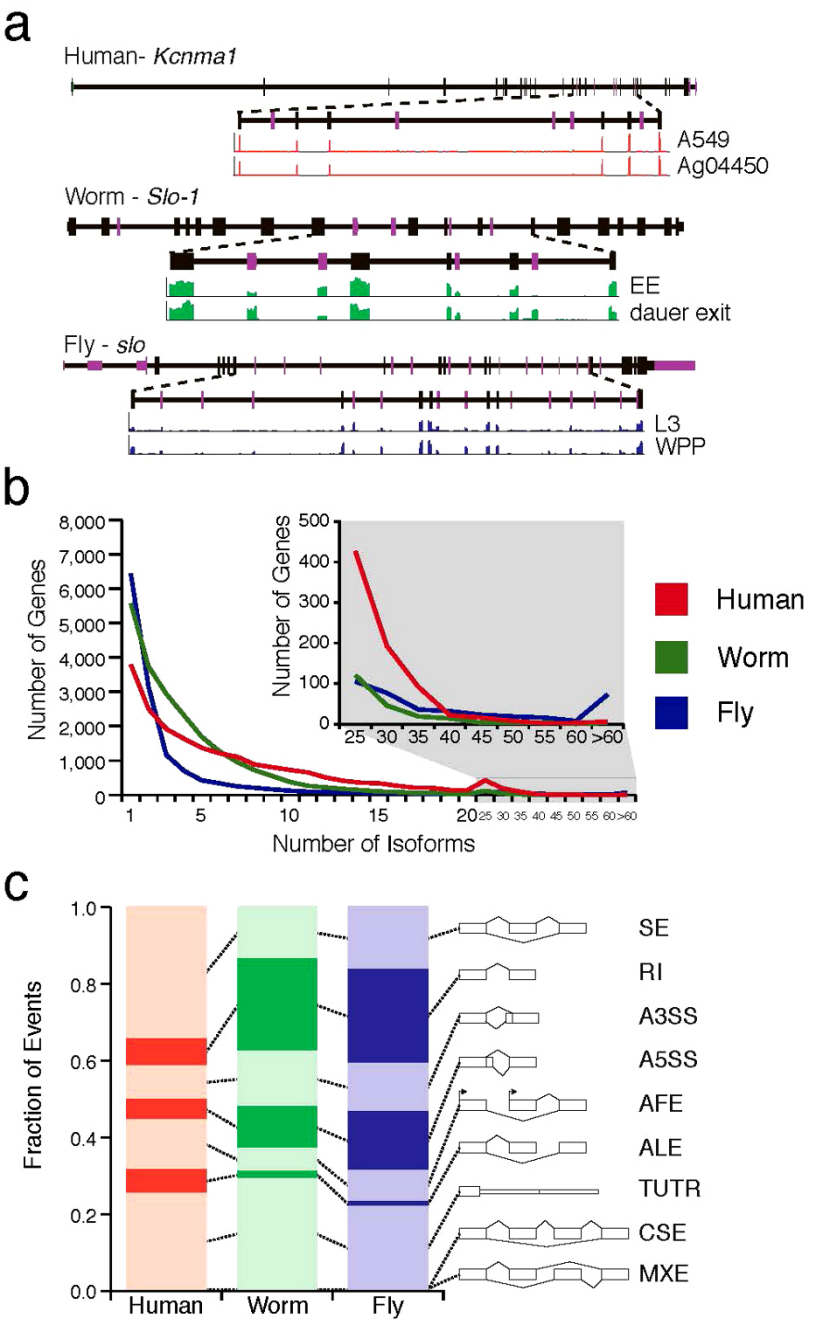


Extended Data Figure 1 | Overview of the data. **a**, Schematic of the RNA-seq data generated for human (red), worm (green) and fly (blue), showing how it samples developmental stages and various tissues and cell lines. **b**, The number and size of data sets generated. The amount of new data beyond that in the previous ENCODE publications[8,11,22] is indicated by white bars, with previous ENCODE data indicated by solid bars. (See Supplementary Information, section B.2, for a detailed description of these data.)

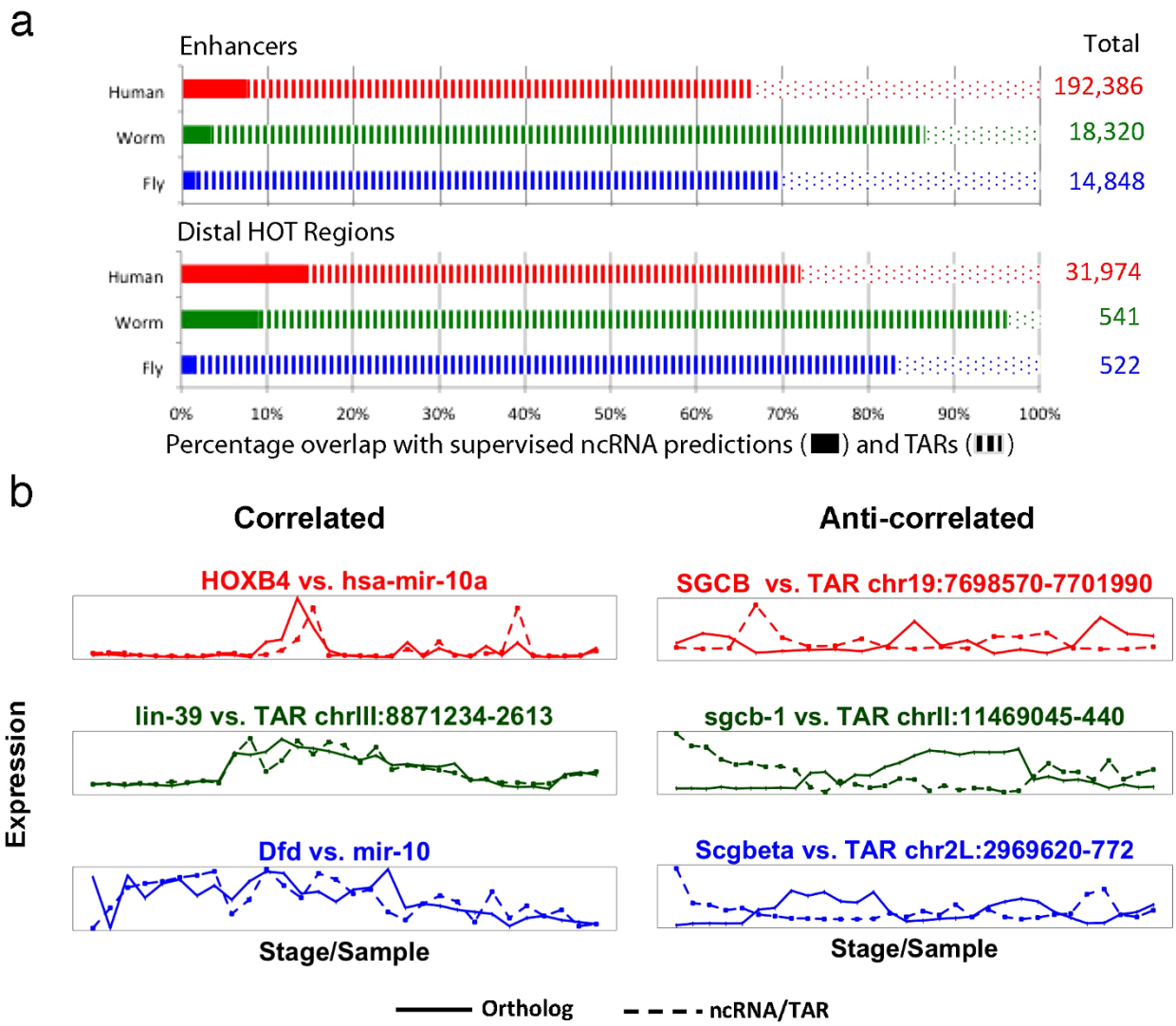
22. Graveley, B. R. *et al.* The developmental transcriptome of *Drosophila melanogaster*. *Nature* **471,** 473–479 (2011).



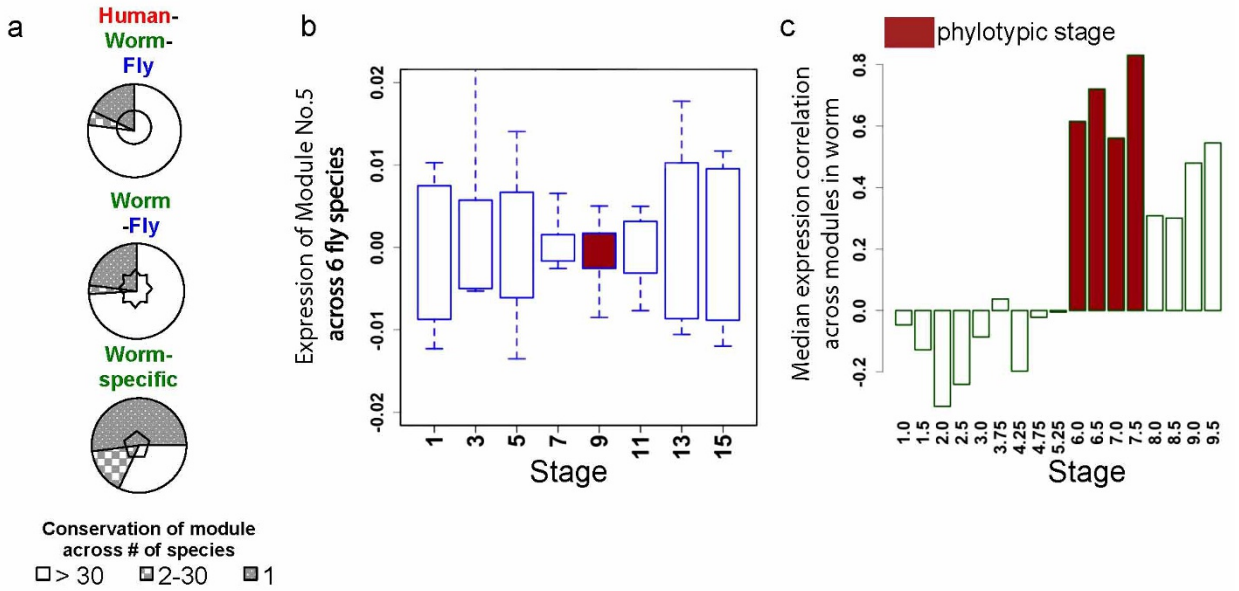
Extended Data Figure 2 | Summary plots for the protein-coding gene annotations. **a**, Distributions of key summary statistics; gene span, longest ORF per gene, CDS exon length, and CDS exons per gene (note that the *x* axes are in log scale). Both fly and worm genes span similar genomic lengths while 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 Supplementary Fig. 1h and Supplementary Information, section B.4.1.)



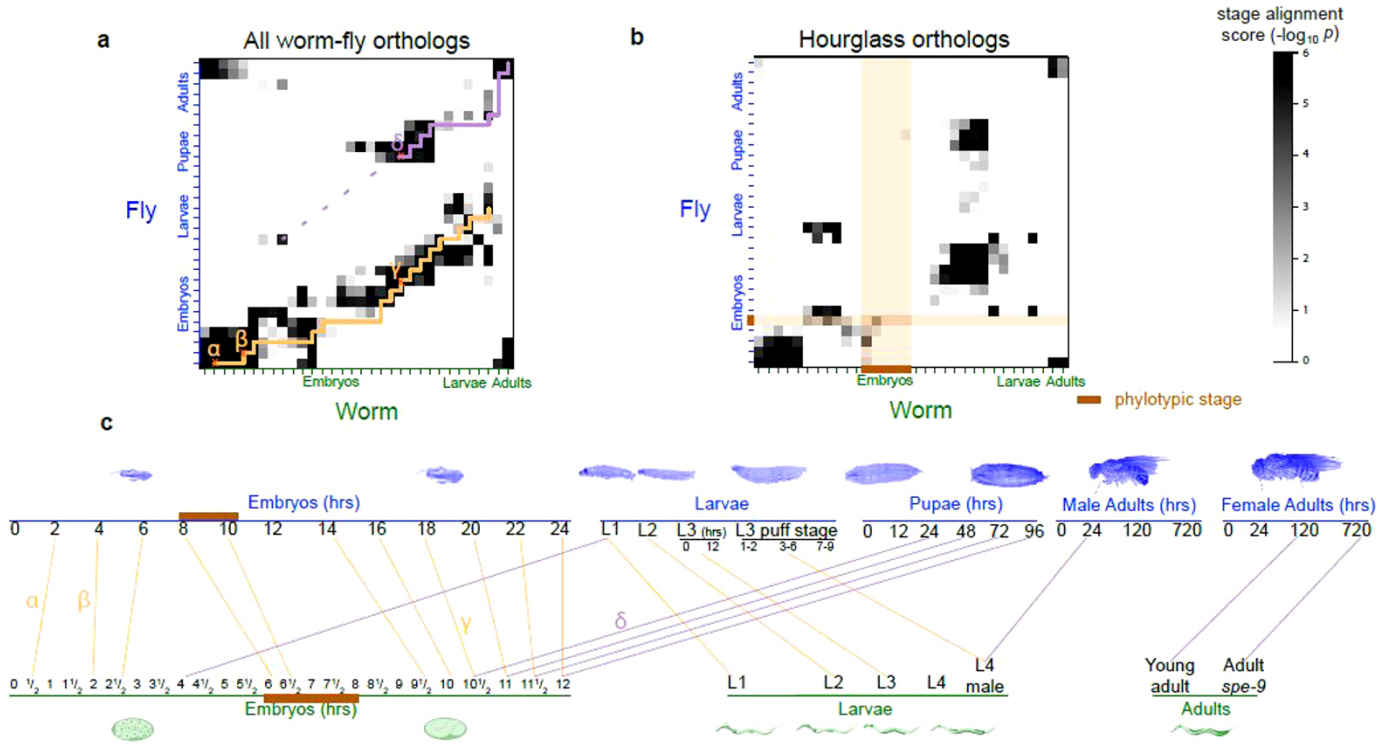
Extended Data Figure 3 | Analysis of alternative splicing. **a**, Representative orthologous genes do not share the same exon-intron structure, or alternative splicing across species. **b**, Distribution of the number of isoforms per gene. **c**, Comparison of the fraction of various alternative splicing event classes in human, worm and fly; A3SS, alternative 3′ splice sites; A5SS, alternative 5′ splice sites; AFE, alternative first exons; ALE, alternative last exons; CSE, coordinately skipped exons; MXE, mutually exclusive exons; RI, retained introns; SE, skipped exons; TandemUTR, tandem 3′ UTRs. (See Supplementary Information, section B.5, for a further discussion of splicing.)



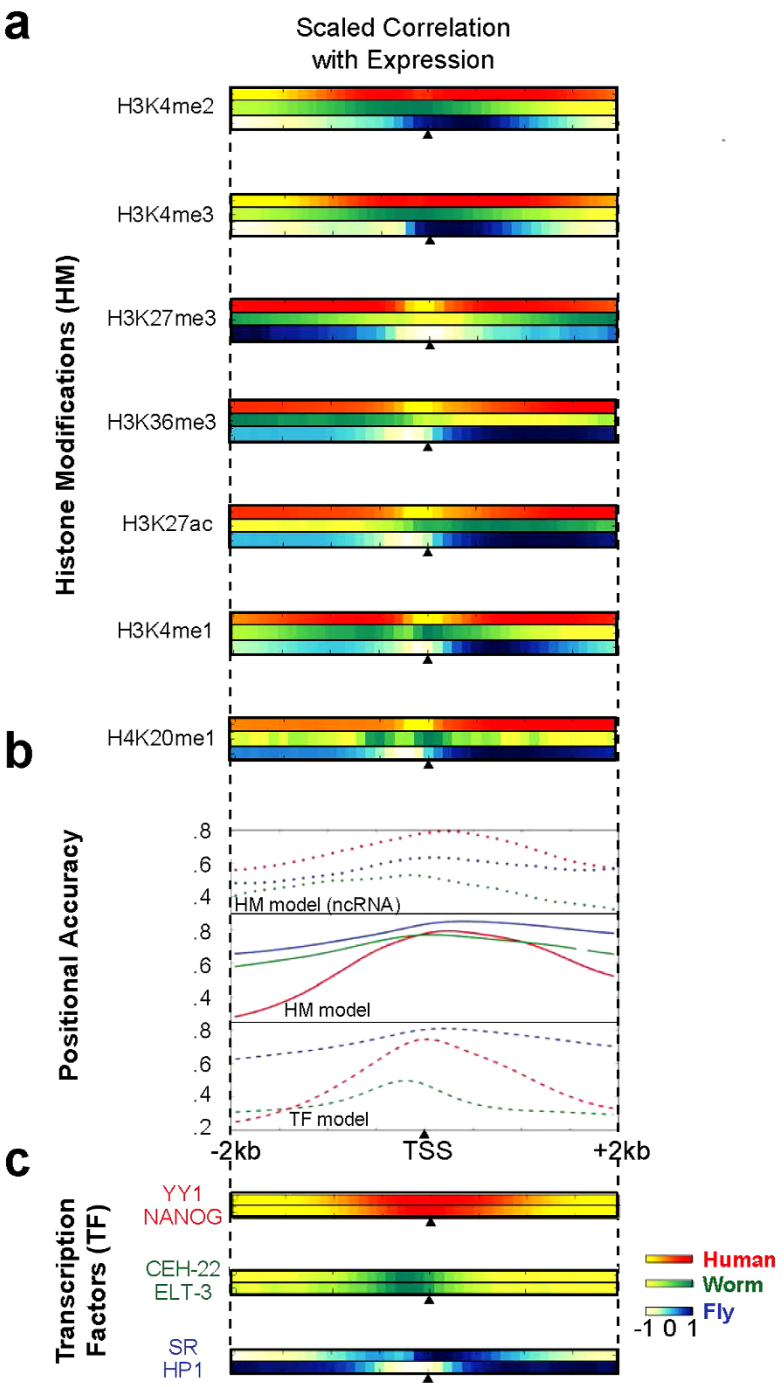
Extended Data Figure 4 | Characterizing non-canonical transcription. **a**, The overlap of enhancers and distal HOT regions with supervised non-coding RNA predictions and TARs in human, worm and fly. The overlap of enhancers and distal HOT regions with respect to both supervised non-coding RNA predictions as well as TARs are significantly enriched compared to a randomized expectation. **b**, The left side highlights non-coding RNA and 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 ($r = 0.66$, $P < 6 \times 10^{-4}$ in fly), as does mir-10a in human, which correlates strongly with HOXB4 ($r = 0.88$, $P < 2 \times 10^{-9}$). A TAR (chr III: 8871234–2613) strongly correlates with *lin-39* ($r = 0.91$, $P < 4 \times 10^{-13}$) in worm. The right side shows TARs in human (chr 19: 7698570–7701990), worm (chr II: 11469045–440), and fly (chr 2L: 2969620–772) that are negatively correlated with the expression of three orthologous genes: SGCB ($r = -0.91$, $P < 3 \times 10^{-16}$), sgcb-1 ($r = -0.86$, $P < 2 \times 10^{-7}$) and Scgb ($r = -0.82$, $P < 4 \times 10^{-8}$), respectively. (More details on all parts of this figure are in Supplementary Information, section C, and Supplementary Table 2.)



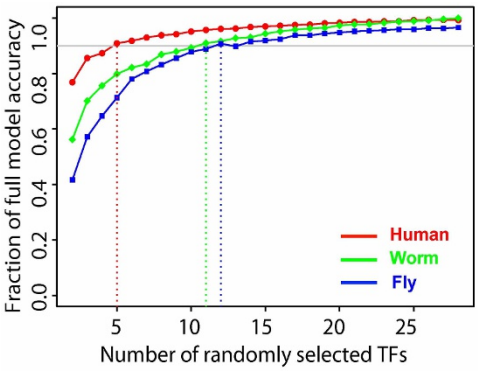
Extended Data Figure 5 | Details on expression clustering. **a**, Pie charts showing gene conservation across 56 Ensembl species for the blocks in the Fig. 1 heatmap enclosed with the same symbol (that is, pentagon here matches pentagon in Fig. 1a). Overall, species-specific modules tend to have fewer orthologues across 56 Ensembl species. **b**, The expression levels of a conserved module (Module No. 5) in *D. melanogaster* and its orthologous counterparts in five other *Drosophila* species are plotted against time. The *x* axis represents the middle time points of 2-h periods at fly embryo stages. The boxes represent the $\log_{10}$ modular expression levels from microarray data of six *Drosophila* species centred by their medians. The modular expression divergence (inter-quartile region) becomes minimal during the fly phylotypic stage (brown, 8–10 h). **c**, The modular expression correlations over a sliding 2-h window (Pearson correlation per five stages, middle time of 2-h period on *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 h. In fact, it is possible to see this coordination directly as a local maximum in the between-module correlation (across time points) for the worm, which has a more densely sampled developmental time course. (This figure provides more detail on Fig. 1a, c. More details on all parts of this figure can be found in Supplementary Information, section D, and Supplementary Fig. 3.)



Extended Data Figure 6 | Details on stage alignment. This figure provides further detail to Fig. 1b. **a**, An alignment of worm and fly developmental stages based on all worm–fly orthologues (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 orthologues. Note the prominent gap in the aligned stages coincides with the worm and fly phylotypic stages (brown band). As the expression values of genes in all hourglass modules converge at the phylotypic stage, no hourglass genes can be phylotypic-stage-specific, thus the gap makes sense. **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 (the primary alignment in Fig. 1b), and they are also matched with fly L1 and prepupa–pupa stages respectively in the 'upper diagonal' set of matches (the secondary alignment in Fig. 1b). (More details on all parts of this figure can be found in Supplementary Information, section D.4, and Supplementary Table 3. See ref. 21 for further details relating to **a** and **c**.)



Extended Data Figure 7 | Further detail on statistical models for predicting gene expression. This figure provides more detailed information than present in Fig. 2. **a**–**c**, Binding or expression correlations of various histone marks (**a**) and transcription factors (**c**). For example, H3K36me3 shows positive correlation in worm and fly, but weak negative correlation in human at the promoter, with positive correlation over the gene body. The positional accuracy from the transcription factor and histone-mark models for predicting mRNA and non-coding RNA expression about the TSS (**b**). (More details on all parts of this figure can be found in Supplementary Information, section E, and Supplementary Fig. 4.)



Extended Data Figure 8 | Average predictive accuracy of models with different number of randomly selected transcription factors. We randomly selected *n* transcription factors as predictors and examined the predictive accuracy by cross-validation, where *n* varied from 2 to 28. The curve shows the average predictive accuracy (Supplementary Fig. 4 indicates the standard deviation of all models with the same number of predictors). Surprisingly, models with as few as five transcription factors 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 transcription facotrs bind somewhat indiscriminately. (More details on all parts of this figure can be found in Supplementary Information, section E.)

Extended Data Table 1 | Summary of annotated non-coding RNAs, TARs and non-coding RNA predictions in each species

			Human Genome Coverage			Worm Genome Coverage			Fly Genome Coverage		
			Elements	Kb	%	Elements	Kb	%	Elements	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

The number of elements, the base pairs covered and the fraction of the genome for each class are shown (see also Supplementary Information, section C). There are comparable numbers of transfer RNAs (tRNAs) in humans and worms but about half as many in fly. Although the number of long non-coding RNAs (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 fivefold more microRNAs (miRNAs), small nucleolar RNAs (snoRNAs) and small nuclear RNAs (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 (249 Mb, 16 Mb and 14 Mb in human, worm and fly, respectively) is intergenic and occurs at low levels, comparable to that observed for introns (Supplementary Table 2). 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 (Supplementary Fig. 2 and Supplementary Information, section C).