LETTER

Comparative analysis of regulatory information and circuits across distant species

Alan P. Boyle[1]*, Carlos L. Araya[1]*, Cathleen Brdlik[1], Philip Cayting[1], Chao Cheng[2], Yong Cheng[1], Kathryn Gardner[3], LaDeana W. Hillier[4], Judith Janette[3], Lixia Jiang[1], Dionna Kasper[3], Trupti Kawli[1], Pouya Kheradpour[6], Anshul Kundaje[5,6], Jingyi Jessica Li[7,8], Lijia Ma[4], Wei Niu[3], E. Jay Rehm[9], Joel Rozowsky[2], Matthew Slattery[9], Rebecca Spokony[9], Robert Terrell[4], Dionne Vafeados[4], Daifeng Wang[2], Peter Weisdepp[4], Yi-Chieh Wu[6], Dan Xie[1], Koon-Kiu Yan[2], Elise A. Feingold[10], Peter J. Good[10], Michael J. Pazin[10], Haiyan Huang[7], Peter J. Bickel[7], Steven E. Brenner[11,12], Valerie Reinke[3], Robert H. Waterston[4], Mark Gerstein[2], Kevin P. White[9]§, Manolis Kellis[6]§ & Michael Snyder[1]§

Despite the large evolutionary distances between metazoan species, they can show remarkable commonalities in their biology, and this has helped to establish fly and worm as model organisms for human biology[1,2]. Although studies of individual elements and factors have explored similarities in gene regulation, a large-scale comparative analysis of basic principles of transcriptional regulatory features is lacking. Here we map the genome-wide binding locations of 165 human, 93 worm and 52 fly transcription regulatory factors, generating a total of 1,019 data sets from diverse cell types, developmental stages, or conditions in the three species, of which 498 (48.9%) are presented here for the first time. We find that structural properties of regulatory networks are remarkably conserved and that orthologous regulatory factor families recognize similar binding motifs *in vivo* and show some similar co-associations. Our results suggest that gene-regulatory properties previously observed for individual factors are general principles of metazoan regulation that are remarkably well-preserved despite extensive functional divergence of individual network connections. The comparative maps of regulatory circuitry provided here will drive an improved understanding of the regulatory underpinnings of model organism biology and how these relate to human biology, development and disease.

Transcription regulatory factors guide the development and cellular activities of all organisms through highly cooperative and dynamic control of gene expression programs. Regulatory factor coding genes are often conserved across deep phylogenies, their DNA-binding protein domains are preferentially conserved at the amino-acid level, and their *in vitro* binding specificities are also frequently conserved across large distances[3,4]. However, the specific DNA targets and binding partners of regulators can evolve much more rapidly than DNA-binding domains, making it unclear whether the *in vivo* binding properties of regulatory factors are conserved across large evolutionary distances.

Comparisons of the locations of regulatory binding across species has been controversial, with some studies suggesting extensive conservation[1,2,5–10], whereas others suggest extensive turnover[11–14]. Although it is generally assumed that across very large evolutionary distances regulatory circuitry is largely diverged, there exist highly conserved sub-networks[15–18]. Thus, confusion exists in the level of regulatory turnover between related species, possibly owing to the small number of factors studied. Moreover, despite recent observations of the architecture of metazoan regulatory networks a direct comparison of their topology and structure—such as clustered binding and regulatory network motif—has not been possible owing to large differences in the procedures employed
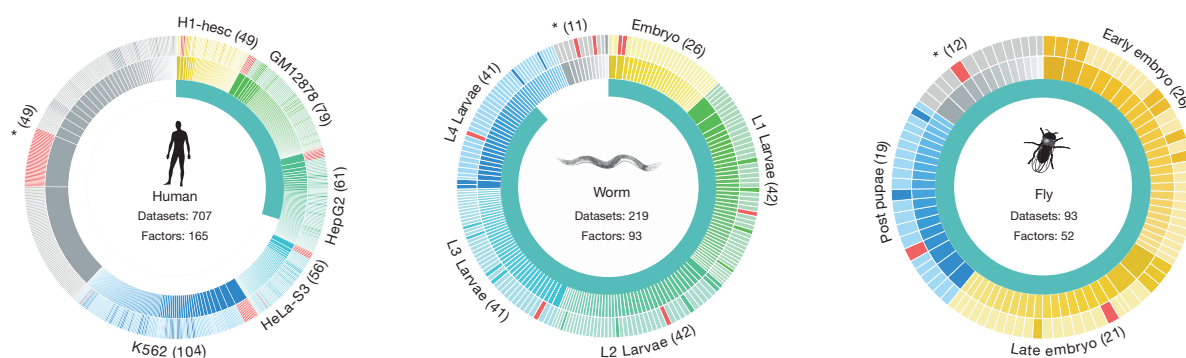

Figure 1 | A large compendium of regulatory binding across distant metazoans. Data from modENCODE and ENCODE consortia used in the analyses. Inner circles show the fraction of data sets presented here for the first time. For each organism the major contexts are shown in a different hue in the two outer circles. Asterisks, data sets that are not one of the main contexts. Each factor that underwent ChIP is shown in the middle circle and the count is in parentheses (a factor can be represented in multiple contexts). The outer circle shows every data set, scaled by the number of peaks. Red, polymerase; light shades, transcription factor; dark shades, other. ChIP of a total of 165, 93 and 52 unique factors across all conditions and cell lines in human, and developmental stages in worm and fly, respectively.

[1]Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA. [2]Program of Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut 06520, USA. [3]Department of Genetics, Yale University School of Medicine, New Haven, Connecticut 06520, USA. [4]Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA. [5]Department of Computer Science, Stanford University, Stanford, California 94305, USA. [6]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. [7]Department of Statistics, University of California, Berkeley, California 94720, USA. [8]Department of Statistics, University of California, Los Angeles, California 90095, USA. [9]Institute for Genomics and Systems Biology, University of Chicago, Chicago, Ilinois 60637, USA. [10]National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, 20892, USA. [11]Department of Molecular and Cell Biology, University of California, Berkeley, California 94720, USA. [12]Department of Plant and Microbial Biology, University of California, Berkeley, California 94720, USA.
*These authors contributed equally to this work.
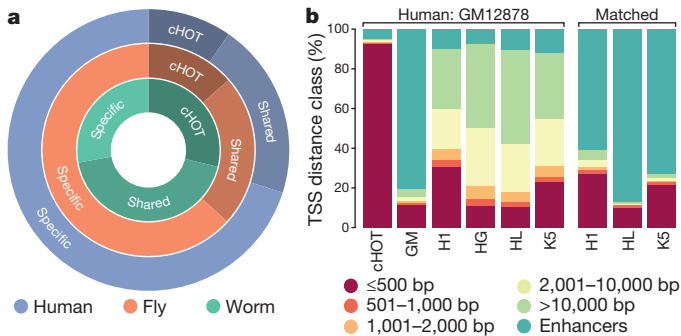§These authors jointly supervised this work.

Figure 2 | HOT regions are dynamically established across stages and cell types. HOT regions contain binding sites for a large number of factors. **a**, A total of 2,948, 2,283, and 46,348 HOT regions exist, of which 29.1%, 13.7% and 9.7% are constitutive in worm, fly and human respectively. A large fraction of HOT regions are shared across multiple contexts but the majority of HOT regions are specific to a single context. **b**, Constitutive human HOT (cHOT) regions show strong enrichment for promoters while cell-type specific (GM12878 (GM), H1hesc (H1), HepG2 (HG), HelaS3 (HL), K562 (K5)) HOT regions show more enhancer enrichment (see also Extended Data Fig. 3). The cell type/context of the classes is indicated on top. Matched indicates that the classes are derived from the specific cell type analysed in each set.

to assay regulatory factor binding in distinct species. Here we present a systematic and uniform comparison of regulation using many factors across distantly related species to help address these questions on a scale not previously possible.

To compare regulatory architecture and binding across diverse organisms, the modENCODE and ENCODE consortia mapped the binding locations of 93 *Caenorhabditis elegans* regulatory factors, 52 *Drosophila melanogaster* regulatory factors and 165 human regulatory factors as a community resource (Fig. 1 and Supplementary Table 1). These regulatory factor binding data sets represent a substantial increase over those previously published for worm (194 new data sets for a total of 219) and

human (211 new, 707 total) and a substantial improvement in data quality in fly with a move from chromatin immunoprecipitation with DNA microarray (ChIP-chip) to ChIP followed by sequencing (ChIP-seq) (93 new, 93 total)[2,8,19,20]. The majority of regulatory factors are site-specific transcription factors (83 in worm, 41 in fly, and 119 in human), although general regulatory factors such as RNA Pol II were also assayed.

All regulatory factors were analysed by ChIP-seq according to mod-ENCODE/ENCODE standards: antibodies were extensively characterized, and at least two independent biological replicates were analysed[21]. Worm regulatory factors were assayed in embryo and stage 1–4 larvae (L1–L4 larvae), fly regulatory factors in early embryo, late embryo and post embryo, and human regulatory factors in myelocytic leukaemia K562 cells, lymphoblastoid GM12878 cells, H1 embryonic stem cells, cervical cancer HeLa cells, and liver epithelium HepG2 cells. Binding sites were scored using a uniform pipeline that identifies reproducible targets using irreproducible discovery rate (IDR) analysis (Extended Data Fig. 1)[22] and quality-filtered experiments (see Methods). These rigorous quality metrics insure that the data sets used here are robust. All data presented are available at http://www.ENCODEProject.org/comparative/regulation/.

To explore motif conservation, we examined the 31 cases in which we had members of orthologous transcription-factor families profiled in at least two species (Extended Data Fig. 2a and Methods). Sequence enriched motifs were found for 18 of the 31 families and for 12 orthologous families (41 regulatory factors), the same motif is enriched in both species (Extended Data Fig. 2b, c). For 18 of 31 families (64 of 93 regulatory factors), the motif from one species is enriched in the bound regions of another species (one-sided hypergeometric, $P = 3.3 \times 10^{-4}$). These findings indicate that many factors retain highly similar *in vivo* sequence specificity within orthologous families, a feature noted previously across studies working on smaller numbers of factors.

Next, we used RNA-seq data[3] to determine whether targets of orthologous regulatory factors are specifically expressed at similar developmental stages between fly and worm. As a class, orthologous regulatory factors (both assayed here and not) are significantly expressed at similar
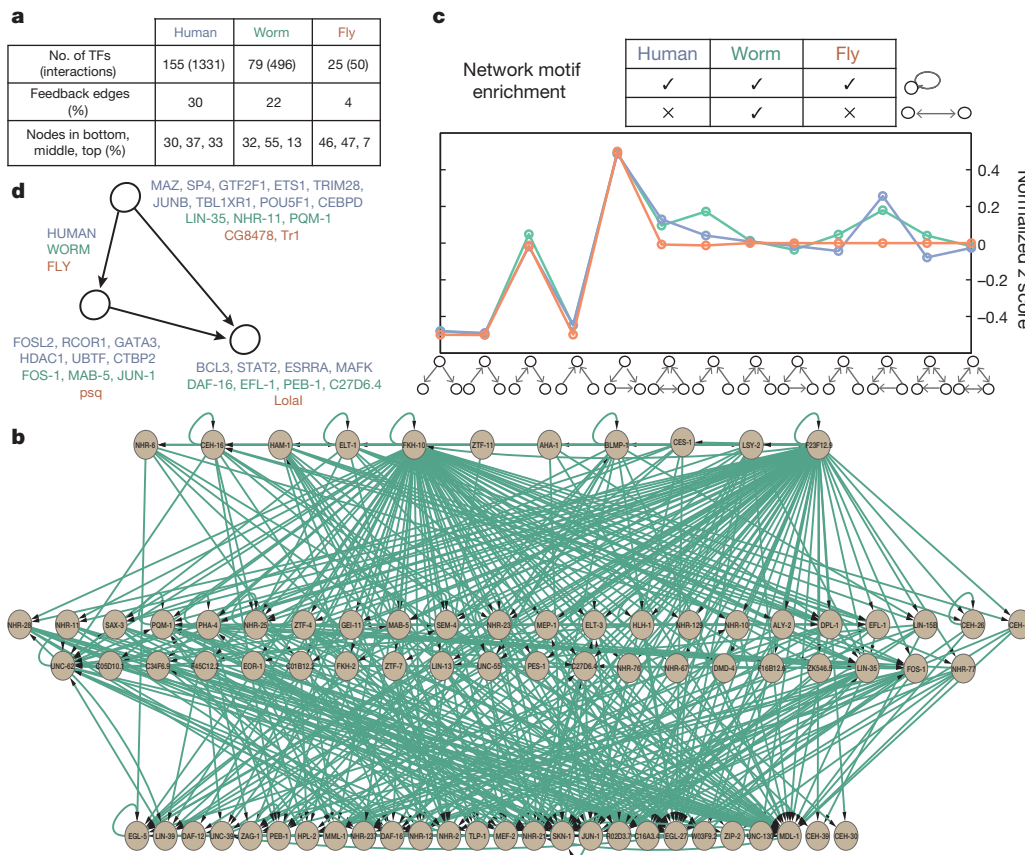

Figure 3 | Regulatory network motif usage is conserved across distant metazoans. a, Statistics of the transcription regulatory networks in human, worm, fly and their hierarchical organization. **b**, An example of the hierarchical network for worm. **c**, Network motif enrichment. The human, worm and fly networks are mostly consistent in terms of motif enrichment. The motif feed-forward loop is the most enriched motif in all three networks. **d**, Different transcription factors have different tendencies to appear as top, middle and bottom regulators in a FFL. The lists of human, worm, fly transcription factors with corresponding tendencies are displayed.

stages (Extended Data Fig. 3a–c). However, expression of orthologous targets of orthologous regulatory factors in worm and fly shows little significant target overlap (Extended Data Fig. 3d) and the large majority of orthologous regulatory factors did not show conserved target functions (Extended Data Fig. 4a–c), suggesting extensive re-wiring of regulatory control across metazoans. Nevertheless, human and worm orthologous regulatory factors were more likely to show conserved target gene functions than non-orthologous regulatory factors (Extended Data Fig. 4d, Wilcoxon test $P < 3.9 \times 10^{-6}$), highlighting regulatory factors with conserved target functions.

Regulatory factor binding is not randomly distributed throughout the genome, but rather, in all three species, approximately 50% of binding events are found in highly-occupied clusters, termed high-occupancy target (HOT) regions[1,2,5,8,10]. HOT regions show enhancer function in integrated transcriptional reporters[11] and are stabilized by cohesin[15,17]. HOT regions show no significant enrichment with non-specific antibodies (Extended Data Fig. 5), in contrast to recent work using raw signal[19] rather than IDR peaks, although the possibility that they are artefacts has been raised.
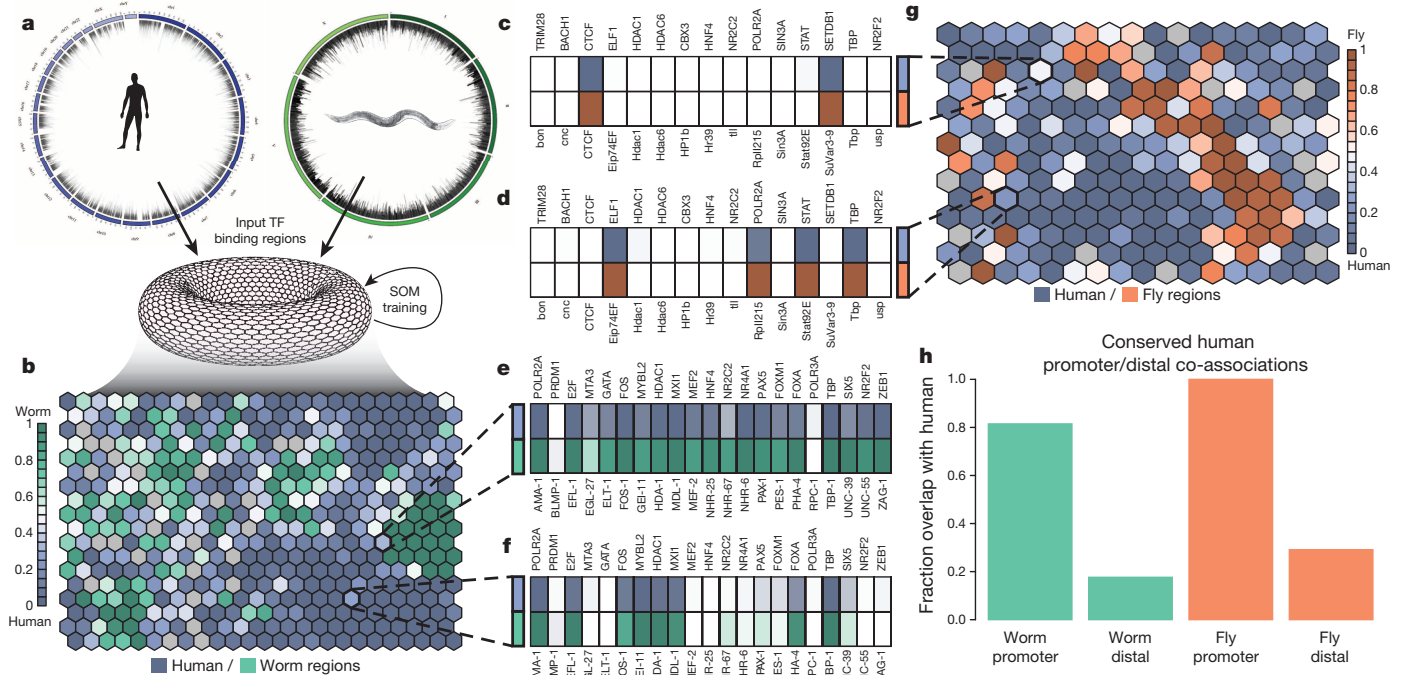
By comparing HOT regions across different developmental times and cells types, we find that 5–10% of HOT regions are constitutive, indicating that HOT regions are dynamically established, rather than an intrinsic property of specific regions. In humans we find that approximately 90% of constitutive HOT regions fall within promoter chromatin states compared to only approximately 10–20% of context-specific HOT regions (Fig. 2a and Extended Data Fig. 6). Instead, approximately 80–90% of context-specific HOT regions fall within enhancer states. Moreover, these context-specific HOT regions are specifically enriched for enhancers in matching cell types or developmental stages. For example, 80% of GM12878-called HOT regions fall within GM12878-specific enhancers but only approximately 10% of GM12878-called HOT regions

fall within enhancers called in other cell-types (Fig. 2b). These patterns remain similar for all cell types (Extended Data Fig. 7), suggesting the two types of HOT regions are established concordantly and dynamically between cell types, though these patterns are weaker in the worm and fly data.

We constructed regulatory networks in each species by predicting gene targets of each regulatory factor using TIP[23] and used simulated annealing to reveal the organization of regulatory factors in three layers of master-regulators, intermediate regulators, and low-level regulators (Fig. 3a, b). The algorithm found only 7% of regulatory factors at the top layer of the network in fly and 13% in worm, compared to 33% in human. We also found that more edges are upward flowing in human (30%) than worm and fly (22% and 7%). This suggests differences in the global network organization with more extensive feedback and a higher number of master regulators in human.

We next assessed the local structure of regulatory networks, by searching for enriched sub-graphs known as network motifs (Fig. 3c). We found that the same network motifs were most and least enriched in the three species. In each case, the most abundant was the feed-forward loop (FFL), while the least abundant were cascade motifs, and both divergent and convergent regulation. Moreover, specific regulatory factors were enriched for origin, target, or intermediate regulators in these FFLs in each species (Fig. 3d). Surprisingly, the number of feed forward loops (FFLs) varied by developmental stage in both worm and fly, with L1 stage in worm and late-embryo stage in fly showing the highest number of FFLs (Extended Data Fig. 8), suggesting increased filtering fluctuations and accelerating responses in these stages[24].

We asked whether the three species showed conserved regulatory factor co-associations. We first focused on global co-associations where two factors co-associate frequently regardless of context, either by intermolecular interactions or independent recruitment (Extended Data

Figure 4 | Transcription-factor co-association. Many instances of transcription-factor co-association are under very specific contexts and probably not observed in a simple genome-wide co-association study. **a**, We combined the patterns of orthologous factors and genomic regions from two organisms to train a SOM where each 'hexagon' contains genomic regions from either organism with the same binding pattern of orthologous factors for worm (**b**) and fly (**g**). Each hexagon is shaded by the frequency of the pattern in the pairs of organisms. We show an example of binding patterns of 4 hexagons from the human–fly (**c–d**) and the human–worm (**e–f**). Names above the heatmaps are human factor names, and those below are their orthologue names. Dark shaded boxes indicate binding of that factor. **c**, A binding pattern shared at equal frequency between human and fly with only CTCF and SETDB1 (CTCF and SuVar3-9 in fly) binding. **d**, A binding pattern that occurs more frequently in human shows ELF1, RNA Pol II, STAT and TBP binding. **e**, A binding pattern at similar frequencies in human and worm that is an example of a HOT region. **f**, A pattern more frequent in humans than worms shows RNA Pol II, E2F, FOS, MYBL2, HDAC1, MXI1, FOXA and TBP binding. **h**, Co-localization patterns that occur more frequently near promoters (<500 bp) in humans are highly likely to also occur at promoters in worm (80%) and fly (100%).

Fig. 9). With the exception of a small number of conserved global regulatory factor co-associations (for example, SIN3A with HDAC1, HDAC2 and NR2C2 in fly and human[25–27], and MXI1 with E2F1, E2F4 and E2F6 in worm and human), the majority of global co-associations were not conserved in the contexts and species pairs analysed.

As regulatory factor co-association at distinct binding regions is local and contextual (that is, different combinations of factors co-associate at different genomic locations), we next used an approach to detect co-association at distinct regions of the genome based on conserved patterns of regulatory factor binding. This method uses self-organizing maps (SOMs) to analyse co-association patterns at specific loci by better exploring the full combinatorial space of regulatory factor binding than traditional co-association approaches (Fig. 4a–c)[28]. We demonstrate that co-associations at distinct genomic regions reveal a more complex view of regulatory structure and bring forth categorical enrichments that are lost in a larger, genomic context.

We examined whether specific contextual co-associations are conserved for orthologous regulatory factors by using binding data from each organismal pair; that is, human–worm and human–fly (Fig. 4b, g). Specific regulatory factor co-associations were observed; most are conserved to varying degrees across each organism with very few that are entirely organism-specific (Fig. 4b, g). These co-associations result in expected sets of factors such as the previously noted SIN3A + HDAC co-association. In addition, we find new co-associations such as the pattern in Fig. 4f for human–worm, which in worm is highly enriched for GO terms associated with sex determination. We further examined which co-associations are conserved at distinct gene locations (that is, proximal and distal). We found distinct combinations of conserved co-associations in relation to transcription start site (TSS) regions. Interestingly, virtually all TSS-proximal co-associations in human remain TSS-proximal in worm (approximately 80%) and fly (approximately 100%), indicating that co-associations that occur at promoters are often highly conserved (Fig. 4h). Conversely, co-associations at distal regions are much less conserved.

Our results, obtained using a large resource of regulatory binding information, suggest that there is little conservation of individual regulatory targets and binding patterns for these highly divergent metazoans: *C. elegans*, *D. melanogaster* and *H. sapiens*. However, we do find strong conservation of overall regulatory architecture, both in network motif usage and in concentrated regulatory binding at dynamically established HOT regions. We observe an increased conservation of *in vivo* sequence preferences and some target gene functions, with context-specific regulatory factor partners still observed at specific loci in these distal comparisons. These findings are consistent with previous results indicating that the gene targets of regulation are typically quite divergent and are likely to account for many of the phenotypic differences among species[12–14,16,29,30], despite conserved sequence preferences. We significantly extend these observations, both in the number of regulators studied and in the range of regulatory properties studied, and provide specific examples of conserved and diverged regulatory functions. Lastly, beyond its potential for comparative studies of gene regulation, the primary data sets provide invaluable new information of genome-wide transcription-factor binding information both in human, and in two of the most important metazoan models of human biology, development, and disease.

1. modENCODE Consortium et al. Identification of functional elements and regulatory circuits by Drosophila modENCODE. *Science* **330,** 1787–1797 (2010).
2. Gerstein, M. B. *et al.* Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science* **330,** 1775–1787 (2010).
3. Gerstein, M. *et al.* Comparative analysis of the transcriptome across distant species. *Nature* http://dx.doi.org/10.1038/nature13424 (this issue).
4. Berger, M. F. *et al.* Variation in homeodomain dna binding revealed by high-resolution analysis of sequence preferences. *Cell* **133,** 1266–1276 (2008).
5. Moorman, C. *et al.* Hotspots of transcription factor colocalization in the genome of *Drosophila melanogaster. Proc. Natl Acad. Sci. USA* **103,** 12027–12032 (2006).
6. Lavoie, H. *et al.* Evolutionary tinkering with conserved components of a transcriptional regulatory network. *PLoS Biol.* **8,** e1000329 (2010).
7. He, Q. *et al.* High conservation of transcription factor binding and evidence for combinatorial regulation across six *Drosophila* species. *Nature Genet.* **43,** 414–420 (2011).
8. ENCODE Project Consortium *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).
9. Mikkelsen, T. S. *et al.* Comparative epigenomic analysis of murine and human adipogenesis. *Cell* **143,** 156–169 (2010).
10. Yip, K. Y. *et al.* Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol.* **13,** R48 (2012).
11. Kvon, E. Z., Stampfel, G., Yáñez-Cuna, J. O., Dickson, B. J. & Stark, A. HOT regions function as patterned developmental enhancers and have a distinct cis-regulatory signature. *Genes Dev.* **26,** 908–913 (2012).
12. Schmidt, D. *et al.* Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328,** 1036–1040 (2010).
13. Odom, D. T. *et al.* Tissue-specific transcriptional regulation has diverged significantly between human and mouse. *Nature Genet.* **39,** 730–732 (2007).
14. Borneman, A. R. *et al.* Divergence of transcription factor binding sites across related yeast species. *Science* **317,** 815–819 (2007).
15. Yan, J. *et al.* Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. *Cell* **154,** 801–813 (2013).
16. Peter, I. S. & Davidson, E. H. Evolution of gene regulatory networks controlling body plan development. *Cell* **144,** 970–985 (2011).
17. Faure, A. J. *et al.* Cohesin regulates tissue-specific expression by stabilizing highly occupied *cis*-regulatory modules. *Genome Res.* **22,** 2163–2175 (2012).
18. Spitz, F. & Furlong, E. E. M. Transcription factors: from enhancer binding to developmental control. *Nature Rev. Genet.* **13,** 613–626 (2012).
19. Teytelman, L., Thurtle, D. M., Rine, J. & van Oudenaarden, A. Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. *Proc. Natl Acad. Sci. USA* **110,** 18602–18607 (2013).
20. Nègre, N. *et al.* A cis-regulatory map of the *Drosophila* genome. *Nature* **471,** 527–531 (2011).
21. Landt, S. G. *et al.* ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* **22,** 1813–1831 (2012).
22. Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.* **5,** 1752–1779 (2011).
23. Cheng, C., Min, R. & Gerstein, M. TIP: a probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles. *Bioinformatics* **27,** 3221–3227 (2011).
24. Alon, U. Network motifs: theory and experimental approaches. *Nature Rev. Genet.* **8,** 450–461 (2007).
25. Heinzel, T. *et al.* A complex containing N-CoR, mSin3 and histone deacetylase mediates transcriptional repression. *Nature* **387,** 43–48 (1997).
26. Nan, X. *et al.* Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. *Nature* **393,** 386–389 (1998).
27. Huang, Y., Myers, S. J. & Dingledine, R. Transcriptional repression by REST: recruitment of Sin3A and histone deacetylase to neuronal genes. *Nature Neurosci.* **2,** 867–872 (1999).
28. Xie, D. *et al.* Dynamic *trans*-acting factor colocalization in human cells. *Cell* **155,** 713–724 (2013).
29. Carroll, S. B., Grenier, J. & Weatherbee, S. *From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design* (Wiley-Blackwell, 2004).
30. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188,** 107–116 (1975).

METHODS

A data portal has been created for the modENCODE project where data from all stages of analysis in this project are available (http://ENCODEProject.org/comparative/regulation/).

Experimental methods for *D. melanogaster* ChIP-seq assay. Transgenic lines containing GFP-tagged transcription factors within their endogenous genomic contexts were produced as described previously[1,31]. Chromatin was collected and chromatin immunoprecipitation was performed as described previously[20]. Multiplexing allowed for sequencing of between 4 and 12 samples per lane on an Illumina Hi-Seq for a minimum of 5 million reads per sample. New GFP-tagged lines are made available at the Bloomington Stock Center. Tagged line stock numbers are: Abd-B stock 38625; Eip74EF stock 38636; Lola stock 38660; N stock 38665; Stat92E stock 38670; usp stock 38672.

Experimental methods for *C. elegans* ChIP-seq assay. *C. elegans* ChIP-seq assays were performed as described in[32], with a few modifications. In brief, transgenic worms containing GFP-tagged transcription factors were grown to the desired developmental stage under controlled conditions and cross-linked with 2% formaldehyde. Cell extracts were sonicated to yield predominantly DNA fragments in the range of 200–500 bp. The sonicated lysates were immunoprecipitated in either 5% or 1% Triton using anti-GFP antibody. Sequencing libraries were prepared from the two independent biological replicates of immunoprecipitation-enriched and input DNA fragments. Libraries were multiplexed using four 4-bp barcodes[33] and sequenced on Illumina Genome Analyzer II.

Experimental methods for human ChIP-seq assay. Human ChIP-seq was performed using the overall method outlined in ref. 21. In brief, 2×10^7 cells were cross-linked using 1% formaldehyde at room temperature followed by treatment with 125 mM glycine. The cross-linked cells were resuspended in hypotonic buffer and the cells were lysed by Dounce homogenization. The resulting nuclear extract was sonicated to obtain DNA fragments in the target size of 200–500 bp. Immunoprecipitation was performed overnight at 4 °C using 2 μg of antibody. The transcription factor–antibody complexes were collected using protein A and Protein G agarose beads. The immunoprecipitation-enriched DNA (transcription-factor antibody as well as control IgG) was used to prepare sequencing libraries similar to the methods used for *C. elegans* ChIP-seq library preparation. A single sample was run per lane of the Illumina Genome Analyzer II.

Uniform processing of transcription factor ChIP-seq data sets. We used a uniform processing pipeline to identify high-confidence binding events (peaks) for a large collection of ChIP-seq data sets in three species from the modENCODE and ENCODE consortia; worm (*C. elegans*), fly (*D. melanogaster*) and human (*H. sapiens*). For human, we analysed 707 distinct ChIP-seq data sets (with at least two replicate experiments) representing 165 unique regulatory factors (generic and sequence-specific factors). The data sets span 91 human cell types and some are in various treatment conditions. These data sets were generated by production groups located at the following universities: The Broad Institute, Stanford University, Yale University, University of California Davis, Harvard University, HudsonAlpha, University of Texas (Austin) and University of Washington. For worm, we analysed 220 distinct ChIP-seq data sets (with at least two replicates) spanning 93 unique regulatory factors in 11 developmental stages. For fly, we analysed 93 distinct ChIP-seq data sets (with at least two replicates each) spanning 52 unique regulatory factors in 17 developmental stages.

Read mapping. For each experiment, mapped reads in the form of BAM files were downloaded from the ENCODE University of California Santa Cruz Data Coordination Center (http://encodeproject.org/ENCODE/downloads.html) and the modENCODE Data Coordination Center (http://www.modencode.org/). These BAM files were generated by the individual data production labs using different mappers and mapping parameters. In order to standardize the mapping protocol, we used custom mappability tracks to filter out multi-mapping reads and only retain unique mapping reads that is, reads that map to exactly one location in the genome. We also filtered all positional and polymerase chain reaction (PCR) duplicates.

Quality control. A number of quality metrics for all replicate experiments of each data set were computed (ref. 21, and A.K., unpublished observations). In brief, these metrics measure ChIP enrichment and signal-to-noise ratios, sequencing depth and library complexity and reproducibility of peak calling. These measures will be reported at the ENCODE portal at http://encodeproject.org/ENCODE/qualityMetrics.html. Data sets that did not pass the minimum quality control thresholds were discarded and not used in any analyses. Data sets that passed most but not all quality metrics were flagged.

Peak calling. All ChIP-seq experiments were scored against an appropriate control designated by the production groups (either input DNA or DNA obtained from a control immunoprecipitation). For human and worm data sets, we used the SPP peak caller to identify and score (rank) potential binding sites and peaks[34]. However, for fly data sets, we instead used the MACS (v.2) peak caller[35]. Most of the fly data sets used the NexTera sample preparation protocol which resulted in non-canonical distribution of reads around binding sites and lower signal to noise ratios. These characteristics made them unsuitable for use with the SPP peak caller which specifically models peak shape and penalizes peaks with non-canonical stranded distribution of reads around binding sites. The MACS v.2 peak caller does not directly model such peak structure and is thus more immune to non-canonical read distributions.

To obtain optimal thresholds, we used the irreproducible discovery rate (IDR) framework to determine high confidence binding events by leveraging the reproducibility and rank consistency of peak identifications across replicate experiments of a data set[22] (A.K., unpublished observations). Code and detailed step-by-step instructions to call peaks using the IDR framework are available at https://sites.google.com/site/anshulkundaje/projects/idr.

For worm and human data sets, the SPP peak caller[34] was used with a relaxed peak calling threshold (FDR = 0.9) to obtain a large number of peaks (maximum of 300,000 for human and 30,000 for worm) that span true signal as well as noise (false identifications). Peaks were ranked using the signal score output from SPP (which is a combination of enrichment over control with a penalty for peak shape). The IDR method analyses a pair of replicates, and considers peaks that are present in both replicates to belong to one of two populations: a reproducible signal group or an irreproducible noise group. Peaks from the reproducible group are expected to show relatively higher ranks (ranked based on signal scores) and stronger rank-consistency across the replicates, relative to peaks in the irreproducible groups. Based on these assumptions, a two-component probabilistic copula-mixture model is used to fit the bivariate peak rank distributions from the pairs of replicates[22].

The method adaptively learns the degree of peak-rank consistency in the signal component and the proportion of peaks belonging to each component. The model can then be used to infer an IDR score for every peak that is found in both replicates. The IDR score of a peak represents the expected probability that the peak belongs to the noise component, and is based on its ranks in the two replicates. Hence, low IDR scores represent high-confidence peaks. An IDR score threshold of 2% for human data sets and 5% for worm data sets was used to obtain an optimal peak rank threshold on the replicate peak sets (cross-replicate threshold). If a data set had more than two replicates, all pairs of replicates were analysed using the IDR method. The maximum peak rank threshold across all pairwise analyses was used as the final cross-replicate peak rank threshold.

Any thresholds based on reproducibility of peak calling between biological replicates are bounded by the quality and enrichment of the worst replicate. Valuable signal is lost in cases for which a data set has one replicate that is significantly worse in data quality than another replicate. Hence, we used a rescue strategy to overcome this issue. In order to balance data quality between a set of replicates, mapped reads were pooled across all replicates of a data set, and then randomly sampled (without replacement) to generate two pseudo-replicates with equal numbers of reads. This sampling strategy tends to transfer signal from stronger replicates to the weaker replicates, thereby balancing cross-replicate data quality and sequencing depth. These pseudo-replicates were then processed using the same IDR pipeline as was used for the true biological replicates to learn a rescue threshold. For data sets with comparable replicates (based on independent measures of data quality), the rescue threshold and cross-replicate thresholds were found to be very similar. However, for data sets with replicates of differing data quality, the rescue thresholds were often higher than the cross-replicate thresholds, and were able to capture more peaks that showed statistically significant and visually compelling ChIP-seq signal in one replicate but not in the other. Ultimately, for each data set, the best of the cross-replicate and rescue thresholds were used to obtain a final rank threshold. Reads from replicate data sets were then pooled and SPP was once again used to call peaks on the pooled data with a relaxed FDR of 0.9. Pooled-data peaks were once again ranked by signal-score. The final rank threshold (best of cross-replicate and rescue threshold) was then used to threshold the ranked set of pooled-data peaks.

For fly data sets, we used a slightly modified version of the above pipeline. For each replicate experiment of a data set, we used the MACS v.2 peak caller[35] with a relaxed P value threshold of 1×10^{-3} to obtain a maximum of 30,000 peaks (replicate sets). Peaks were ranked based on their P values. Reads from the replicate experiments were then pooled and once again MACS v.2 was used with a P-value threshold of 1×10^{-3} to obtain a relaxed set of peaks (pooled set). We only retained peaks in the pooled set that overlapped at least one peak in both replicate sets (replicate-reproducible peaks). For each replicate-reproducible peak in the pooled set, we obtained a pair of P values corresponding to the overlapping peaks in each of the replicate sets. If a peak in the replicate-reproducible set overlapped multiple peaks in a replicate-set then the P value of the replicate-set peak with the maximal overlap with the pooled-set peak was used. Thus, we obtain two independent rankings based on P values from each replicate for the same set of replicate-reproducible peaks (using peak coordinates learned on the pooled set). The pair of ranked lists for the replicate-reproducible peaks were then used as input to the IDR framework as

described above to learn cross-replicate rank thresholds at an IDR of 5%. The above protocol was repeated for pseudo-replicates to obtain a rescue rank threshold at an IDR of 5%. The better of the two rank thresholds was used to truncate the replicate-reproducible peaks in the pooled set to obtain the final set of optimal rank consistent and reproducible peaks.

All peak sets were then screened against specially curated empirical blacklists for each species. In brief, these blacklist regions typically show the following characteristics: unstructured and extreme high signal in sequenced input-DNA and control data sets as well as open chromatin data sets irrespective of cell-type identity; an extreme ratio of multi-mapping to unique mapping reads from sequencing experiments; overlap with specific types of repeat regions such as centromeric, telomeric and satellite repeats that often have few unique mappable locations interspersed in repeats.

Identification of HOT and XOT regions. To identify regions with higher than expected binding occupancies, we first determined for each specific context in each organism the number and size distribution of observed binding sites for each factor assayed, as well as the total number and size distribution of binding regions in which these binding sites from all factors are clustered. For each target case (context per species evaluated), we first analysed the number and size distribution of target binding regions (in which factor binding sites are concentrated). For each target case simulation, we randomly selected an equivalent number of random binding regions with a matched size distribution. Next, for each factor assayed (in the target case), we evaluated the number and size of observed binding sites, and simulated an equivalent number and size distribution of target binding sites, restricting their placement to the simulated binding regions. We collapsed simulated binding sites from all factors into binding regions, verifying that these cluster into a similar number of simulated binding regions as the target binding regions. For each target case simulation, the occupancy (number of peaks), density (peaks per kb), and complexity (diversity of factors) in the simulated binding regions are annotated. This procedure was repeated 1,000 times for each case (human = 5 contexts; worm = 5 contexts; fly = 3 contexts). For each target case, we constructed expected binding region occupancy distributions from the corresponding 1,000 simulations. We determined the cutoffs at which fewer than 5% and 1% of the simulated binding regions have higher occupancies (Extended Data Fig. 2). We classified observed binding regions with occupancies higher than the 5% and 1% cutoffs as high-occupancy target (HOT) and extreme-occupancy target (XOT) regions, respectively. As such, HOT regions include XOT regions.

GO enrichment analysis. To evaluate the functional role of regulators we performed GO enrichment analysis on the targets of binding of each ChIP-seq experiment. In brief, we applied ChIPpeakAnno to assign factor binding to genic targets and to evaluate the enrichment of genic targets for GO ontologies using standard procedures[36]. We required a minimum of 20 peaks per ChIP-seq experiment to evaluate enrichment and report Benjamini–Hochberg corrected P values of enrichment (hypergeometric testing). We report GO terms in which at least one ChIP-seq experiment was significantly enriched (corrected $P < 0.05$). The specific enrichments for each human, worm and fly ChIP-seq experiment are provided in Supplementary Tables 2, 3, and 4, respectively.

To compare the functional conservation of regulatory binding between transcription-factor orthologues, we evaluated the overlap in GO term enrichments for orthologous factors between species. Specifically, for each species comparison, we calculated the significance of the overlap in GO term enrichments for all ChIP-seq experiments involving orthologous factors assayed in the two species. Overlap enrichment and depletion P values between ChIP-seq experiments of each species were determined using directional Fisher's exact tests and were Benjamini–Hochberg-corrected. To generate a final overlap score, we selected the most significant of the enrichment and depletion scores, reporting the $-\log_{10}(P$ value of enrichment) or the $\log_{10}(P$ value of depletion).

Generation of orthologue list. Analysis was performed on twelve *Drosophila* species (*D. melanogaster, D. simulans, D. sechellia, D. yakuba, D. erecta, D. ananassae, D. pseudoobscura, D. persimilis, D. willistoni, D. mojavenis, D. virilis, D. grimshawi*) using the September 2010 release of FlyBase, five *Caenorhabditis* species (*C. elegans, C. brenneri, C. briggsae, C. japonica, C. remanei*) using WormBase WS220, and two mammals (*H. sapiens, Mus musculus*) and one out-group species (*Saccharomyces cerevisiae*) using Ensembl release 61.

Gene families were defined using Ensembl Compara gene families for the primary species (human, mouse, *D. melanogaster, C. elegans, S. cerevisiae*), and these clusters were supplemented by genes from the additional fly and worm species using BLAST[37]. For each gene family, we aligned the peptide sequences using MUSCLE[38]. From this alignment, we built an initial gene tree using RAxML[39] with the PROTGAMMAJTT model, then corrected for topological uncertainty using TreeFix[40], and finally accounted for possible incomplete lineage sorting using DLCoal[41]. For DLCoal, we used species tree parameters from literature for the main species and assumed that the remaining fly and worm species take the parameters of

D. melanogaster or *C. elegans*, respectively. To infer homologues, we considered two genes as orthologues (paralogues) if their most recent common ancestor is a speciation (duplication) node. To improve orthologue calls, we filtered out duplications was zero consistency score[42]. Finally, we remapped Ensembl identifiers to release 65.

From a total of 31,751 identified gene families within the three genomes, our data sets here capture 242, with 34 families having a transcription factor from at least two species (100 transcription factors, 459 data sets), and 6 families from all three species (24 transcription factors, 130 data sets). Overall, we found 14 pairs of homologous factors between worm and fly (corresponding to 12 transcription factors in worm and 12 in fly), 41 pairs between worm and human (23 and 36 transcription factors, respectively) and 28 pairs between fly and human (17 and 24 transcription factors, respectively). 14 orthologous triplets were in common for all three organisms (corresponding to 10 transcription factors in human, 8 in fly, and 6 in worm).

Many of these factors are quite divergent in sequence among the species with the exception of RNA polymerases II and III, histone deacetylases, and TBP. Multiple experiments in different stages were available for many factors and some of the common factors are expressed at analogous times for worm and fly[3].

Motif enrichment. We restrict our analysis to the entire genome excluding HOT regions, unmappable and blacklist regions, 3′ untranslated regions (UTRs), coding exons, and several other exons for human (ribosomal RNAs, small nucleolar RNAs, and other miscellaneous RNAs, small nuclear RNAs, microRNAs) and worm (all available). These background regions are randomly split into two groups: one for motif discovery and another for ranking the enrichment of the discovered motifs. Transcription factors that have more than five available data sets in a species have five data sets randomly selected for analysis. Motif discovery is conducted on the top 200 peaks for each data set that overlap the discovery background using five discovery tools: AlignACE[43] (v.4.0 with default parameters), MDscan[44] (v.2004 with default parameters), MEME[45] (v.4.7.0 with -maxw 26 and -nmotifs 6), Weeder[46] (v.1.4.2 with option large), and Trawler[47] (v.1.2 with 200 random intergenic blocks for background). For each species and factor family, the top three motifs are selected after ranking by the enrichment in the data sets for that species and excluding motifs for which a similar motif has already been selected (Pearson $r > 0.7$). These discovered motifs are augmented with all known literature motifs for factors in that gene family[48–50].

Enrichments are computed by taking the fraction of motif instances that are inside the bound regions and dividing that by the fraction of shuffle motif instances inside (where the bound regions are filtered against the background regions, defined below). They are also corrected for small counts by using a Wilson's binomial confidence interval (with $Z = 1.5$) around each fraction and taking the extreme which leads to the enrichment closest to 1. Motifs are considered enriched if this corrected enrichment is at least 1.5-fold.

The discovered motifs, their enrichments, and the underlying annotations are available at http://www.broadinstitute.org/~pouyak/motif-disc/integrate-cold/.

Enrichment of orthologous transcription-factor expression. To match the developmental stages of *D. melanogaster* and *C. elegans*, we first estimated the expression levels of orthologous genes between fly and worm at different developmental stages by applying Cufflinks[51] to modENCODE time course RNA-seq data. We next identified stage-associated genes—genes highly expressed at that stage but not always highly expressed across all stages—for every fly and worm developmental stage. Then for every possible pair of fly and worm stages, we counted the number of orthologous gene pairs between their stage-associated genes, which would be used to test against the null hypothesis that the fly and worm stages have independent stage-associated genes. For the resulting p values, we applied Bonferroni correction and used the corrected P values to decide which fly and worm stages 'match' (have dependent stage-associated genes).

Transcription-factor co-association (intervalStats). We determined the similarity in binding sites between ChIP-seq experiments applying recently developed interval statistics methods that allow calculation of exact P values for proximity between binding sites[52]. Using this method, we performed all pairwise comparisons of ChIP-seq experiments for each organism, evaluating binding similarity in 114,582 human comparisons, 34,782 worm comparisons, and 3,906 fly comparisons. For each species, we restrained interval analyses to the promoter domains by excluding binding intervals outside promoter regions. To exclude the possibility of promiscuous binding regions and generate more conservative co-association estimates, we excluded binding sites from XOT regions in each specific context from these analyses. Promoter regions were defined as 5,000 bp upstream to 500 bp downstream of human TSSs, and 2,000 bp to 200 bp downstream of worm and fly TSSs. Focusing co-association analyses on the promoter domains serves to focus co-association evaluations on transcriptional regulatory interactions and to account for the known biases in binding at TSSs and produces more conservative estimates of co-association significance. For each comparison, the intervals of the query ChIP-seq experiment are compared individually against all reference intervals of the alternate

ChIP-seq experiment, calculating the probability that a randomly located query interval of the same length would be at least as close to the reference set. For each comparison, we compute the fraction of proximal binding events in promoter domains that are significant (P value <0.05). Because these comparisons are asymmetric—depending on the assignment of experiments as query or reference sets—we report the mean values of the complementary (inverted) comparisons.

Transcription-factor co-association (SOM). Using the orthologous factors between human-worm or human-fly, we defined a cis-regulatory module as the maximum overlapping block of the intersection of all transcription-factor binding peaks on either genome. We require a minimum of two transcription factors bound in a *cis*-regulatory module to be considered for further analysis in the self-organizing map (SOM). Several window sizes were examined for co-association (500 bp, 1 kb, and DNase hypersensitive sites[53]) with similar results found in each case.

We binarized each *cis*-regulatory module as either bound (1) or not bound (0) by overlap with peaks from each transcription factor. This results in the *cis*-regulatory modules being represented as a binary vector of the number of dimensions being the count of orthologous transcription-factor families. These vectors, which map back to specific genomic locations, are now directly comparable across species. These are used as input to the SOM and resulting descriptions of each neuron are also described in this form.

For each SOM trained, we followed the rules described previously in ref. 28. In brief, these rules are: the SOM is initialized as a random toroid; the SOM is hexagonal; the SOM is trained for 100 epochs (that is, complete iterations through the data set); the SOM update radius was one-third of the map size with a learning rate (alpha) of 0.05 (these were linearly decreased throughout the training process); the best out of 1,000 trials, based on lowest quantization error, were selected for analysis (defined as the average Euclidean distance of all CRMs to their best matching neuron).
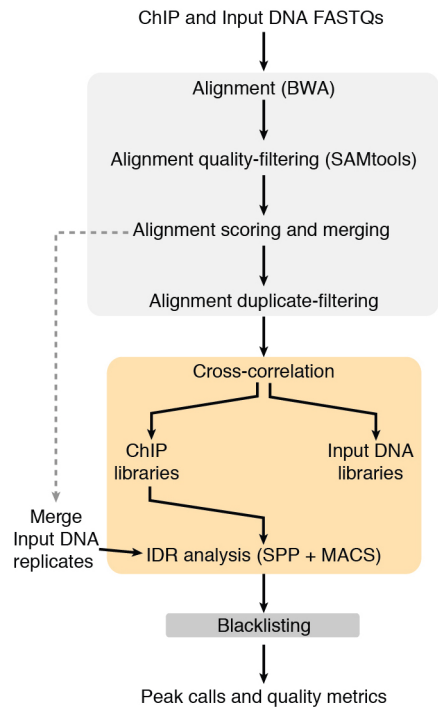
The training described above is performed in R using a variant of the 'kohonen' package available from CRAN[54]. Minor modifications were performed to the R package to allow for better handling of the large data sets in memory. Furthermore, significant changes to the graphical output of the package were made to allow for the improved figures displayed here and on the supplementary website. Final optimal seeds for the training were human–worm SOM: 49,027 and human–fly SOM: 60938. One hundred epochs of training resulted in stabilization of the classification error, and of the 1,000 iterations of the SOM there was minimal divergence with the best SOM having less than 0.3% difference in error than the average error of the non-optimal SOMs. Final SOM sizes were 25×18 and 17×14 for the human–worm and human–fly SOMs respectively and average CRM distance to the best matching neuron was 0.429 and 0.308 for human–worm and human–fly respectively.
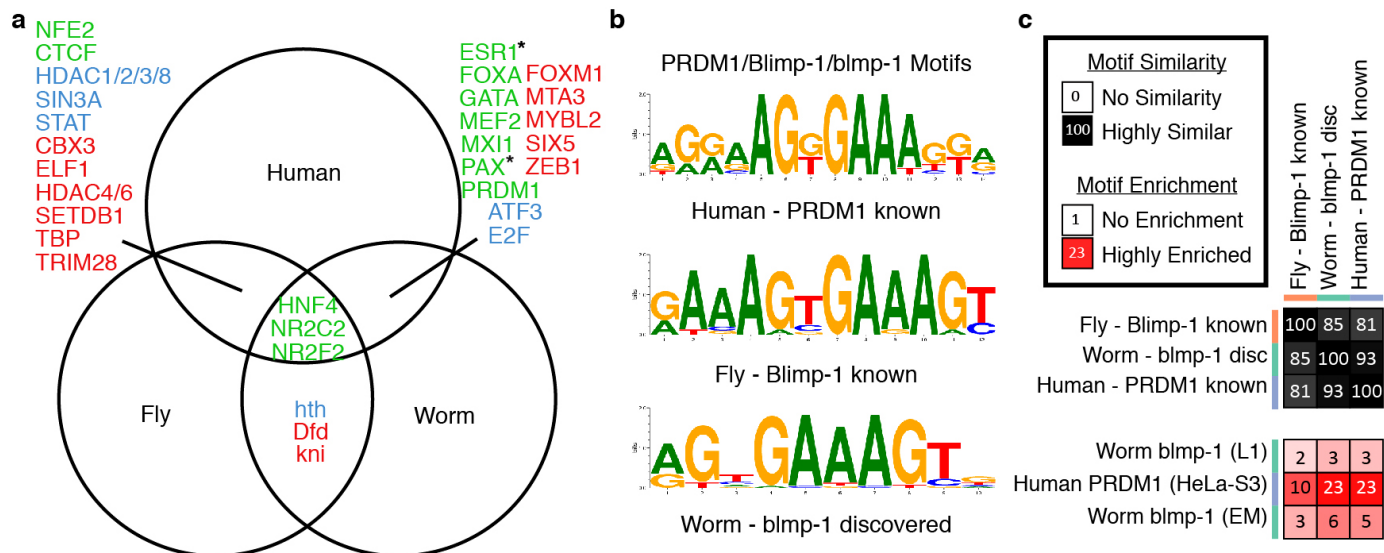
Interactive SOMs can be accessed at http://ENCODEProject.org/comparative/regulation/Worm/SOM/ and http://ENCODEProject.org/comparative/regulation/Fly/SOM/.

Regulatory-network construction. The targets of individual transcription factors in human, worm and fly were identified using TIP[23]. The regulatory networks are the superposition of all the regulatory edges in the three species respectively. For the analysis of transcription factor–transcription factor regulatory networks (Fig. 3a, c, d), we used a Q-value threshold of 0.1 in all three species. For the analysis including various target genes, a Q-value threshold of 0.01 was employed. In Fig. 3a, b, the hierarchical organization was constructed by assigning the nodes in three levels such that an energy function based on the number of feedback edges was minimized. For enrichment analysis (Fig. 3c, d) the null model is an ensemble of random networks with the same degree distribution as the network of interest. In part **d**, the tendency of a transcription factor at a particular position of a FFL is obtained by counting how often it appears at the position in the network of interest, and how often in appears at the same position in the null model.

31. Venken, K. J. T. *et al.* Versatile P[acman] BAC libraries for transgenesis studies in *Drosophila melanogaster. Nature Methods* **6,** 431–434 (2009).
32. Zhong, M. *et al.* Genome-wide identification of binding sites defines distinct functions for *Caenorhabditis elegans* PHA-4/FOXA in development and environmental response. *PLoS Genet.* **6,** e1000848 (2010).
33. Lefrançois, P. *et al.* Efficient yeast ChIP-Seq using multiplex short-read DNA sequencing. *BMC Genomics* **10,** 37 (2009).
34. Kharchenko, P. V., Tolstorukov, M. Y. & Park, P. J. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature Biotechnol.* **26,** 1351–1359 (2008).
35. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9,** R137 (2008).
36. Zhu, L. J. *et al.* ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics* **11,** 237 (2010).
37. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).
38. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32,** 1792–1797 (2004).
39. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22,** 2688–2690 (2006).
40. Wu, Y.-C., Rasmussen, M. D., Bansal, M. S. & Kellis, M. TreeFix: statistically informed gene tree error correction using species trees. *Syst. Biol.* **62,** 110–120 (2013).
41. Rasmussen, M. D. & Kellis, M. Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome Res.* **22,** 755–765 (2012).
42. Vilella, A. J. *et al.* EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* **19,** 327–335 (2009).
43. Hughes, J. D., Estep, P. W., Tavazoie, S. & Church, G. M. Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae. J. Mol. Biol.* **296,** 1205–1214 (2000).
44. Liu, X. S., Brutlag, D. L. & Liu, J. S. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nature Biotechnol.* **20,** 835–839 (2002).
45. Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2,** 28–36 (1994).
46. Pavesi, G. *et al.* MoD Tools: regulatory motif discovery in nucleotide sequences from co-regulated or homologous genes. *Nucleic Acids Res.* **34,** W566–W570 (2006).
47. Ettwiller, L., Paten, B., Ramialison, M., Birney, E. & Wittbrodt, J. Trawler: *de novo* regulatory motif discovery pipeline for chromatin immunoprecipitation. *Nature Methods* **4,** 563–565 (2007).
48. Matys, V. *et al.* TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34,** D108–D110 (2006).
49. Bryne, J. C. *et al.* JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.* **36,** D102–D106 (2008).
50. Newburger, D. E. & Bulyk, M. L. UniPROBE: an online database of protein binding microarray data on protein–DNA interactions. *Nucleic Acids Res.* **37,** D77–D82 (2009).
51. Roberts, A., Pimentel, H., Trapnell, C. & Pachter, L. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* **27,** 2325–2329 (2011).
52. Li, J. J., Huang, H., Bickel, P. J. & Brenner, S. E. Comparison of *D. melanogaster* and *C. elegans* developmental stages by modENCODE RNA-Seq data. *Genome Res.* **24,** 1086–1101 (2014).
53. Boyle, A. P. *et al.* High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132,** 311–322 (2008).
54. Wehrens, R. & Buydens, L. M. Self-and super-organizing maps in R: the Kohonen package. *J. Stat. Softw.* **21,** 1–19 (2007).
55. Chikina, M. D. & Troyanskaya, O. G. An effective statistical evaluation of ChIPseq dataset similarity. *Bioinformatics* **28,** 607–613 (2012).
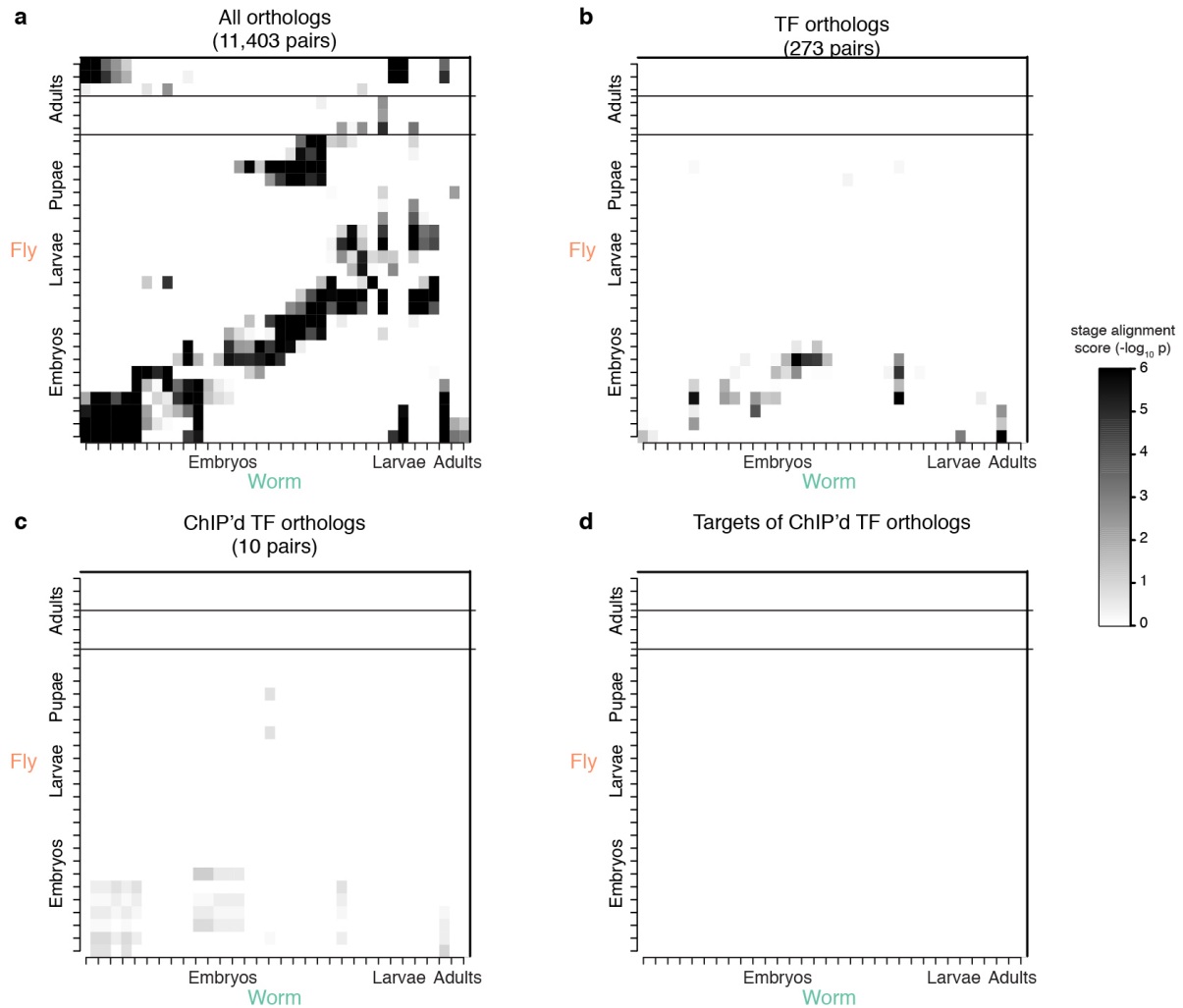
ChIP and Input DNA FASTQs

↓

Alignment (BWA)

↓

Alignment quality-filtering (SAMtools)

┄┄┄ Alignment scoring and merging

↓

Alignment duplicate-filtering

↓

Cross-correlation

ChIP libraries   Input DNA libraries

Merge Input DNA replicates → IDR analysis (SPP + MACS)

↓

Blacklisting

↓

Peak calls and quality metrics

Extended Data Figure 1 | Outline of data-processing pipeline. All data sets were processed using a uniform processing pipeline with identical alignment and filtering criteria and standardized IDR peak calling using SPP (human + worm) and MACS2 (fly).

a

NFE2
CTCF
HDAC1/2/3/8
SIN3A
STAT
CBX3
ELF1
HDAC4/6
SETDB1
TBP
TRIM28

Human

ESR1*
FOXA   FOXM1
GATA   MTA3
MEF2   MYBL2
MXI1   SIX5
PAX*   ZEB1
PRDM1
ATF3
E2F

HNF4
NR2C2
NR2F2

Fly

hth
Dfd
kni

Worm

b

PRDM1/Blimp-1/blmp-1 Motifs

Human - PRDM1 known

Fly - Blimp-1 known

Worm - blmp-1 discovered

c

Motif Similarity

| 0 | No Similarity |
| 100 | Highly Similar |

Motif Enrichment

| 1 | No Enrichment |
| 23 | Highly Enriched |

	Fly - Blimp-1 known	Worm - blmp-1 disc	Human - PRDM1 known
Fly - Blimp-1 known	100	85	81
Worm - blmp-1 disc	85	100	93
Human - PRDM1 known	81	93	100

	Fly - Blimp-1 known	Worm - blmp-1 disc	Human - PRDM1 known
Worm blmp-1 (L1)	2	3	3
Human PRDM1 (HeLa-S3)	10	23	23
Worm blmp-1 (EM)	3	6	5

Extended Data Figure 2 | Motifs. a, Thirty-two transcription-factor gene families with a binding data set for at least two species (names abbreviated). Cross enrichment indicates the enrichment of motifs from one species in the data sets of another. For 13 families, we observed no cross enrichment (red). For 7 families (blue) we observed cross enrichment and for an additional 12 (green) we also had matching motifs. For two cases marked by an asterisk a known fly motif matches the human motif but no worm motif matches. **b**, PRDM1 (also known as Blimp-1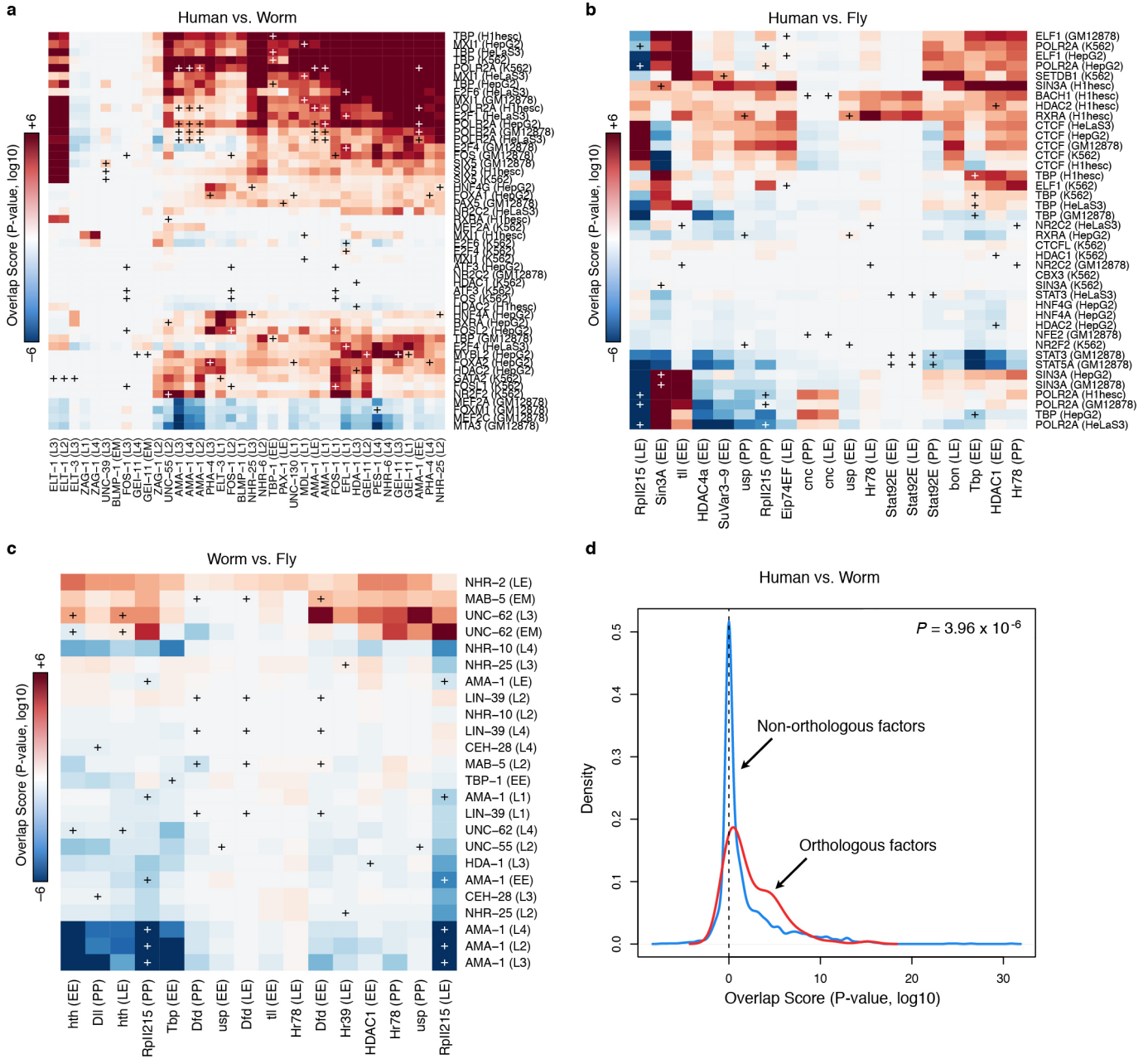 in worm) gene family. We discovered a motif in worm data sets that matches literature-derived known motifs from human and fly. **c**, All three motifs are highly similar and enriched in human PRDM1 and worm blmp-1 data sets. Cell-type and treatment are indicated for each data set in parenthesis. Enrichments in each box are the fraction of motif instances that are inside the bound regions and dividing that by the fraction of shuffled motif instances. Additional motifs known and discovered for these and other data sets are included in Supplementary Information.
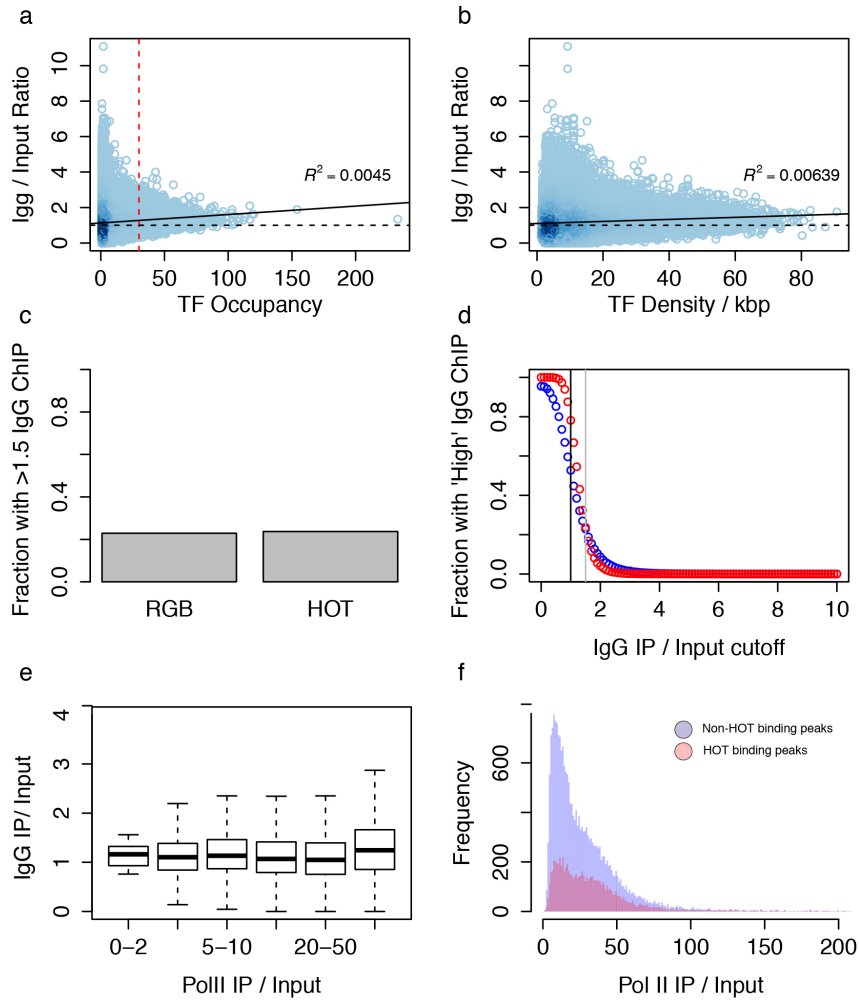
a, Fly–worm stage alignment of expression using all fly–worm orthologues.
b, Alignment of fly–worm stage using all transcription-factor orthologues.
c, Alignment of fly–worm stage using transcription-factor orthologue that has undergone ChIP. **d**, Alignment of fly–worm stage using proximal genes to transcription-factor binding sites that has undergone ChIP. The stage-mapped data exhibit two sets of collinear patterns between the two species (distinct diagonals). In the bottom diagonal, expression from worm embryos and larvae are matched with fly embryos and larvae, respectively. Worm adults are

matched with fly early embryos and fly female adults, possibly owing to the orthologous gene expression in eggs of both species; worm dauers are matched with fly late embryo to L1 and L3 stages, which is similar to the position of dauer stages in the worm lifecycle (between worm L1 and L4 stages). In the upper diagonal, worm middle embryos are matched with fly L1 stage; worm late embryos are matched with fly prepupae and pupae stages; worm L4 male larvae are matched with fly male adults. This collinear pattern may be attributable to fly genes with two-mode expression profiles and many-to-one fly-worm orthologous gene pairs. For more details, please refer to the companion paper[55].

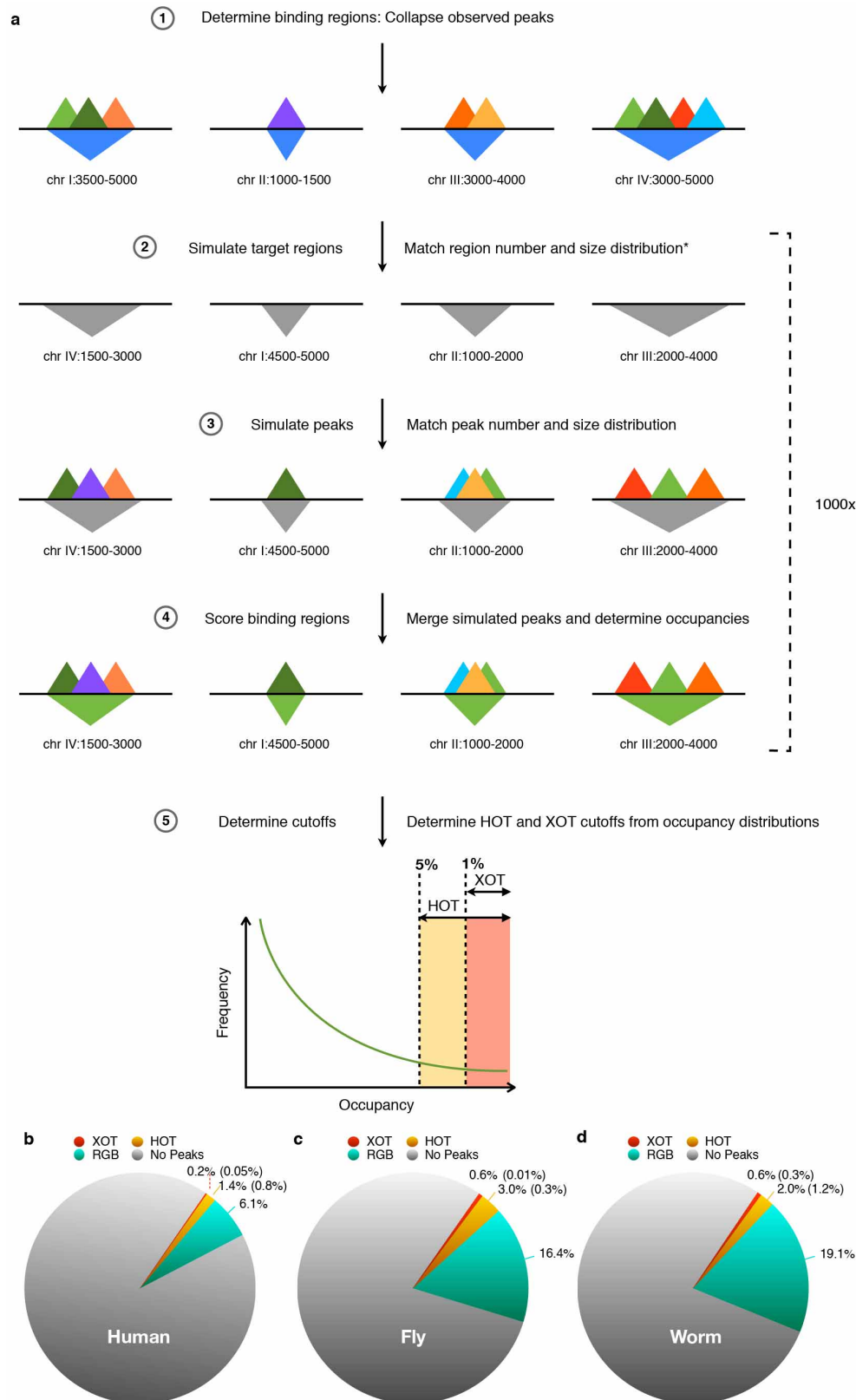**Extended Data Figure 3 | Orthologous expression in worm and fly.**

Extended Data Figure 4 | Comparison of GO enrichment of orthologous transcription-factor pairs. A comparison of GO enrichment of orthologous transcription-factor pairs for all contexts in human versus worm (**a**), human versus fly (**b**), and worm versus fly (**c**) is shown. Red boxes indicate level of similar GO enrichment. 'Plus' signs mark orthologous transcription-factor pairs with white 'pluses' indicating the most significant enrichment for an orthologue pair. **d**, Orthologous factors are more enriched for matching GO terms than non-orthologous factors.

Extended Data Figure 5 | Human HOT enrichments are not overly enriched for control DNA. HOT regions do not represent assembly or ChIP-ability artefacts. **a**, Scatter plot of IgG immunoprecipitation or input versus transcription-factor occupancy. Scatterplot is shaded by density of points. Red dash line represents HOT threshold and black dashed line represent a 1× enrichment. Black line represents the line of best fit for the scatter plot ($R^2 = 0.0045$). **b**, A scatterplot of density (number of transcription-factor peaks per kb) rather than total number of peaks in a region shows a similar trend. **c**, Barplot of fraction of regions with high IgG enrichment for HOT and non-HOT (RGB) regions using the same threshold (1.5×), as ref. 19 revealed little similarity between HOT regions and artefact ChIP regions. **d**, The fraction of HOT (red) and non-HOT (blue) regions with high IgG enrichment is plotted as a function of threshold. Black line represents no enrichment (IgG to input ratio = 1×) and grey dashed line represents the enrichment cutoff (1.5×) used in **b** and in Fig. 7 of ref. 19. **e**, Comparison of IgG (IgG to input ratio) and RNA Pol II enrichment (RNA PolII to input ratio) shows a different trend from Fig. 3a of ref. 19. **e**, Nearly all (99.967%) of our uniformly processed RNA Pol II binding sites have immunoprecipitation to input ratios of greater than 2×, with a median enrichment of approximately 20×.
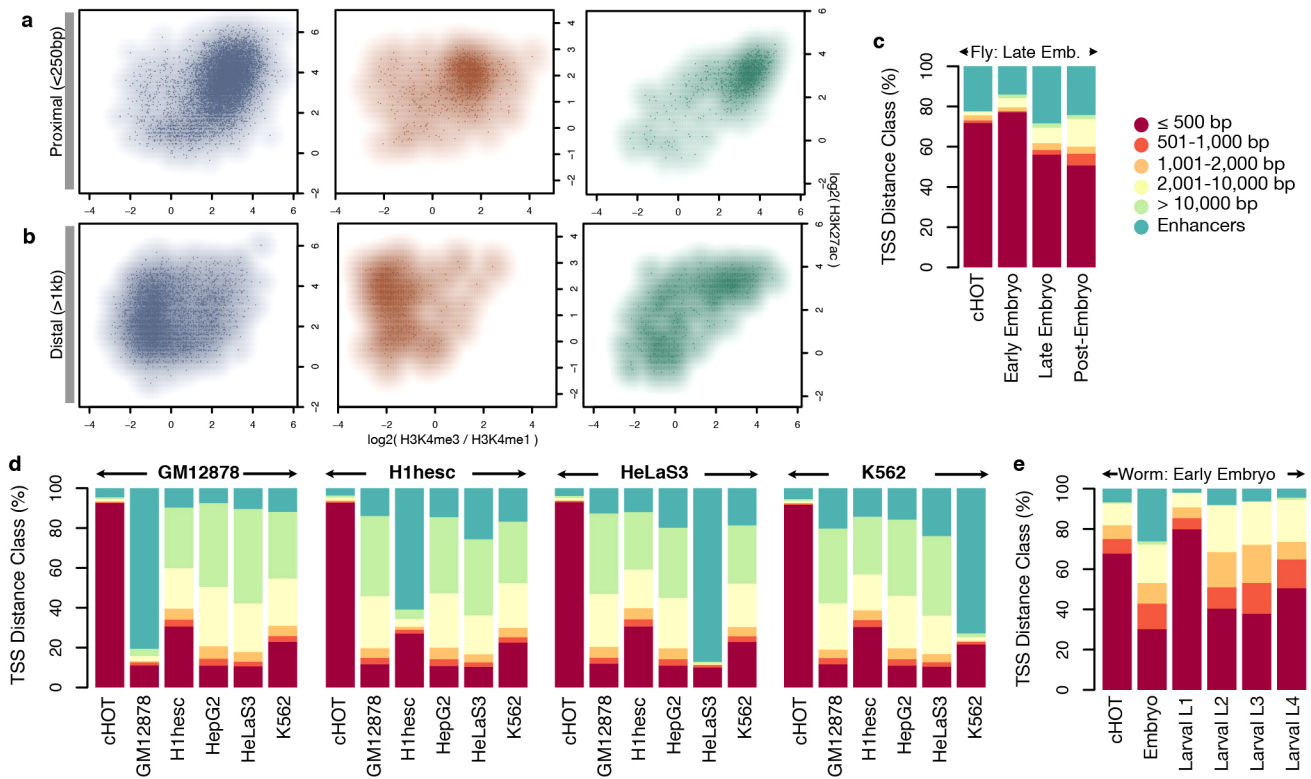
Extended Data Figure 6 | HOT regions were identified in all organisms.
a, To identify HOT region for each context, we first analysed the number and size distribution of target binding regions (in which factor binding sites are concentrated). For each target case simulation, we randomly select an equivalent number of random binding regions with a matched size distribution. Next, for each factor assayed (in the target case), we evaluated the number and size of observed binding sites, and simulated an equivalent number and size distribution of target binding sites, restricting t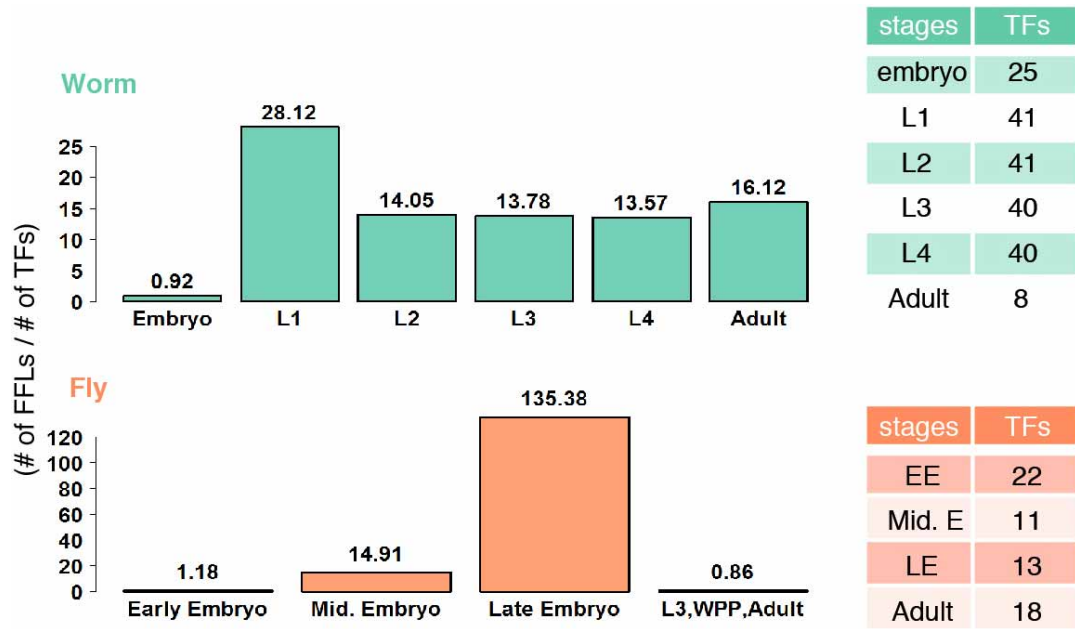heir placement to the simulated binding regions. We collapsed simulated binding sites from all factors into binding regions, verifying that these cluster into a similar number of simulated binding regions as the target binding regions. We identify regions at a 5% (HOT) and 1% (XOT) occupancy threshold based on this simulated data. **b**, Binding of regulatory factors covers different fractions of the genomes of fly, human, and worm. Coverage is shown for constitutive HOT (cHOT, red), HOT (yellow), and non-HOT (RGB, green) regions. Coverage for XOT regions is given in brackets.
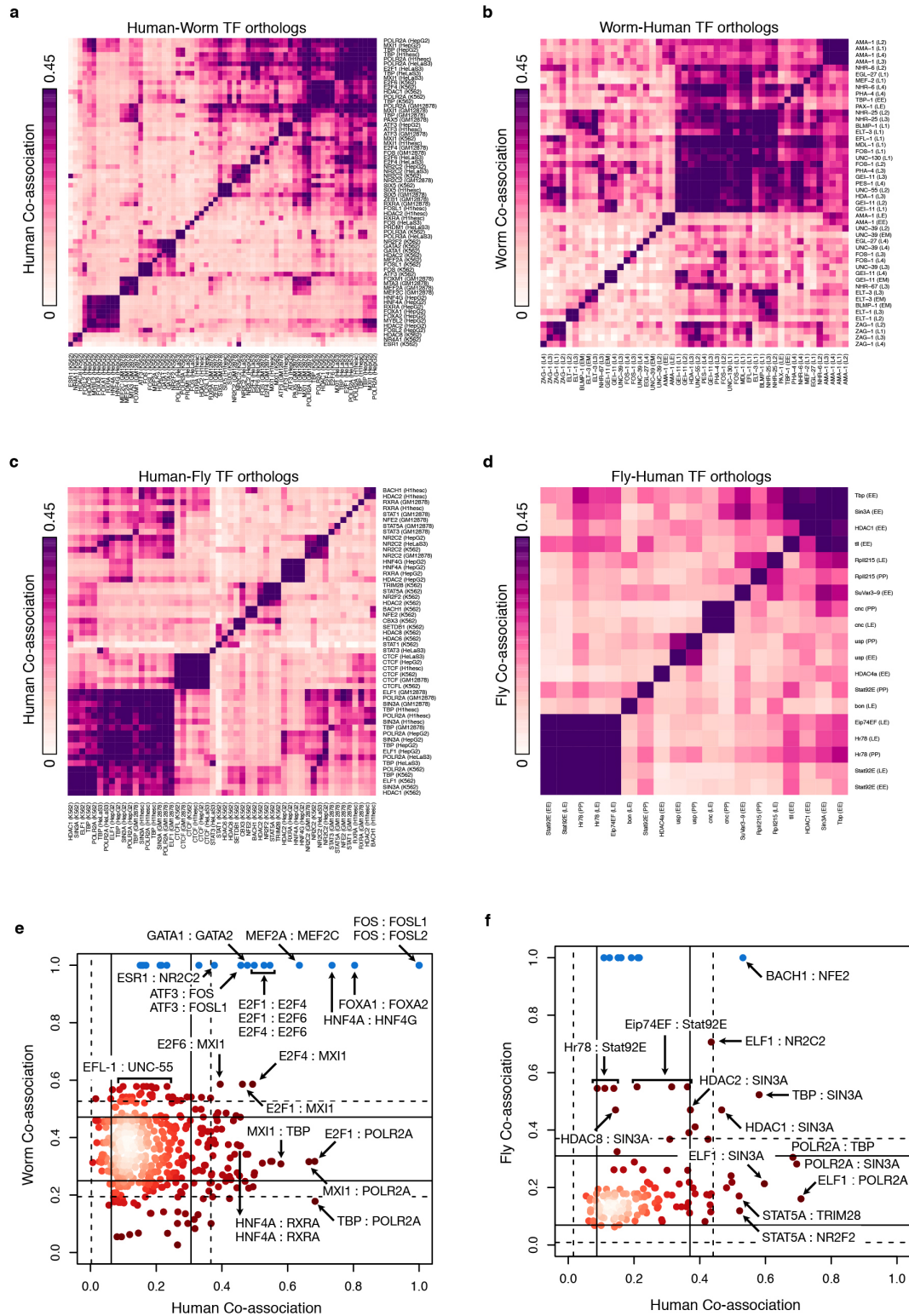
Extended Data Figure 7 | HOT enrichments with context-specific enhancer enrichments. a, **b**, Histone marks for HOT regions (represented by points and smoothed to show density) at proximal (**a**) and distal sites (**b**) show similar trends of histone mark enrichment in their flanking regions. Enhancer calls for a specific developmental stage (**c**, **e**) or cell type (**d**) (labelled over each set of bar graphs) match HOT regions from that cell type and not HOT regions from another cell type. Each set of six bar graphs represents the same set of HOT regions called constitutively HOT or specific to each of the five cell types. Constitutive HOT (cHOT) regions are significantly enriched at promoters with the remaining regions overlapping enhancer regions.

Extended Data Figure 8 | The number of feed forward loops in different stage-specific networks. The number of FFLs in a stage is normalized by the number of transcription factors in the corresponding stage-specific network. Although the sets of transcription factors may differ, the number of transcription factors in each stage stays roughly the same.

Extended Data Figure 9 | Co-associations. Evolutionary retention and change in transcription-factor co-associations. The pairwise co-association strengths between orthologous transcription factors are shown for human–worm orthologues (**a**, **b**) and human–fly orthologues (**c**, **d**). For each pair of species-specific orthologues across multiple samples, the co-association strength, measured as the fraction of significant co-binding events between experiments, is shown (IntervalStats[52]). **a**, Human co-association matrix for human–worm orthologues. **b**, Worm co-association matrix for human-worm orthologues. **c**, Human co-association matrix for human–fly orthologues. **d**, Fly co-association matrix for human–fly orthologues. **e**, Comparison of human–worm transcription-factor orthologue co-associations. The co-association

strength of human–worm orthologues in human (*x* axis) is plotted against the co-association strength in worm (*y* axis). Lines depict 1 (solid) and 1.5 (dashed) standard deviations from the mean score. Factors in blue represent enrichments due to paralogous transcription factors in human that tend to be highly co-associated. **f**, Comparison of human–fly transcription-factor orthologue co-associations. Co-association strength in human (*x* axis) is plotted against co-association strength in fly (*y* axis). For transcription-factor orthologues assayed in multiple developmental stages and/or cell lines, the maximal co-association between contexts was selected for the comparative analyses (**e**, **f**).