nature genetics

ARTICLES

https://doi.org/10.1038/s41588-022-01089-w

Check for updates

OPEN

Multiomic atlas with functional stratification and developmental dynamics of zebrafish *cis*-regulatory elements

Damir Baranasic[1,2,35], Matthias Hörtenhuber[3,35], Piotr J. Balwierz[1,2,4,35], Tobias Zehnder[1,2,5,35], Abdul Kadir Mukarram[3,35], Chirag Nepal[6], Csilla Várnai[4,7], Yavor Hadzhiev[4], Ada Jimenez-Gonzalez[4], Nan Li[4], Joseph Wragg[4], Fabio M. D'Orazio[4], Dorde Relic[8], Mikhail Pachkov[8], Noelia Díaz[9,33], Benjamín Hernández-Rodríguez[9], Zelin Chen[10,11,12], Marcus Stoiber[13], Michaël Dong[3], Irene Stevens[3], Samuel E. Ross[14], Anne Eagle[15], Ryan Martin[15], Oluwapelumi Obasaju[4], Sepand Rastegar[16], Alison C. McGarvey[17], Wolfgang Kopp[17], Emily Chambers[18], Dennis Wang[18,19], Hyejeong R. Kim[20], Rafael D. Acemel[21,22], Silvia Naranjo[21], Maciej Łapiński[23], Vanessa Chong[24], Sinnakaruppan Mathavan[25], Bernard Peers[26], Tatjana Sauka-Spengler[24], Martin Vingron[5], Piero Carninci[27,34], Uwe Ohler[17,28], Scott Allen Lacadie[17], Shawn M. Burgess[11], Cecilia Winata[23], Freek van Eeden[20], Juan M. Vaquerizas[1,2,9], José Luis Gómez-Skarmeta[21], Daria Onichtchouk[29], Ben James Brown[13], Ozren Bogdanovic[14,30], Erik van Nimwegen[8], Monte Westerfield[15], Fiona C. Wardle[31], Carsten O. Daub[3,32] ✉, Boris Lenhard[1,2] ✉ and Ferenc Müller[4] ✉

Zebrafish, a popular organism for studying embryonic development and for modeling human diseases, has so far lacked a systematic functional annotation program akin to those in other animal models. To address this, we formed the international DANIO-CODE consortium and created a central repository to store and process zebrafish developmental functional genomic data. Our data coordination center (https://danio-code.zfin.org) combines a total of 1,802 sets of unpublished and re-analyzed published genomic data, which we used to improve existing annotations and show its utility in experimental design. We identified over 140,000 *cis*-regulatory elements throughout development, including classes with distinct features dependent on their activity in time and space. We delineated the distinct distance topology and chromatin features between regulatory elements active during zygotic genome activation and those active during organogenesis. Finally, we matched regulatory elements and epigenomic landscapes between zebrafish and mouse and predicted functional relationships between them beyond sequence similarity, thus extending the utility of zebrafish developmental genomics to mammals.

Zebrafish is used as a model vertebrate in over 1,200 laboratories worldwide for studies of organismal, cell and gene function in development, regeneration, behavior, toxicology and disease modeling. Its popularity is due to its experimental advantages[1], convenient genetic manipulation tools, wide-ranging genetics resources (for example, Zebrafish Information Network; ZFIN[2]), and high conservation of disease genes and mechanisms between human and fish[3]. Use of zebrafish in genomic studies has discovered chromatin signatures[4–6], DNA codes of promoter usage[7], regulatory patterns of DNA methylation and post-transcriptional messenger RNA regulation[8–13], while zebrafish single-cell genomics pioneered applications for spatially resolving lineage-specific transcriptomes during development[14], and comparative genomics has predicted conserved regulatory elements and their long-range target genes[15]. Exploiting the ease of zebrafish transgenesis, automated in vivo imaging and image processing, which can be upscaled to high throughput[16], provided validation of predicted disease-associated human enhancers[17,18]. However, despite these many landmark studies, zebrafish has lacked systematic functional annotation programs at a scale seen in other models, including ENCODE[19], Roadmap Epigenome[20,21] and modENCODE[22,23]. Thus, disparate zebrafish genomic datasets remain largely inaccessible to thousands of laboratories. Furthermore, while promoters and enhancers from some adult zebrafish tissues have been annotated[24], embryonic and larval stages lack functional annotation despite representing the bulk of zebrafish-based research. Recognizing these needs, DANIO-CODE was established as a multinational bottom-up effort[25].

DANIO-CODE aimed to functionally annotate the developing zebrafish genome by (1) collecting all published and producing new genomic data from 38 laboratories worldwide and standardizing metadata annotation; (2) creating and maintaining a single data coordination center (DCC) for continued accumulation and user download of zebrafish genomic datasets[26]; (3) developing standardized analysis pipelines and remapping all sequencing datasets; and

A full list of affiliations appears at the end of the paper.

(4) generating an integrated track hub that allows visualization with common genome browsers. Additionally, DANIO-CODE aimed to conduct an integrated analysis of these datasets to promote discovery, functional element classification and determination of features of developmental dynamics. Finally, in this study, we applied new approaches for comparative analysis of zebrafish and mammalian genomic datasets to uncover conservation of the genomic landscape and to expand the utility of zebrafish developmental genomics resources.

Results

The DANIO-CODE DCC. We established a DCC protocol[26], which we populated with zebrafish developmental genomic data, including standardized annotation of metadata of diverse, often inconsistently annotated, published datasets (Fig. 1a), by the DANIO-CODE consortium (https://www.birmingham.ac.uk/generic/danio-code/partners/index.aspx). The DCC is accessible from ZFIN and includes datasets, their underlying samples and sequencing protocols using ZFIN and ENCODE nomenclature (www.danio-code.zfin.org). To identify and analyze the developmental dynamics of genomic features, direct comparison across datasets produced by different laboratories and different protocols is required. To this end, we carried out consistent reprocessing starting from the raw sequencing data (Fig. 1a). Raw sequencing data were collected and reprocessed by standardized pipelines of ENCODE for ChIP–seq and ATAC-seq[27], FANTOM for CAGE-seq[28] and producer pipelines for Hi-C, 4C-seq or other data (Methods). These pipelines are available on GitLab (https://gitlab.com/danio-code). The DCC data include 1,438 published datasets contributed by data producers directly or collected by DANIO-CODE data annotators, together with strategically selected datasets for developmental stages from the public domain. In addition, 366 datasets were generated by consortium members to fill gaps and to aid functional annotation and functional element characterization, including 15 CAGE-seq, 18 ChIP–seq, 11 ATAC-seq, 2 Hi-C and 320 4C-seq datasets (Fig. 1b and Extended Data Fig. 1a,b). Breakdown of the datasets according to data types and stages of development is presented in Fig. 1b. The source of data collection is in Extended Data Figure 1c and Supplementary Table 1. Quality checks and data comparability analyses were carried out for datasets within a data type obtained from multiple laboratories, particularly affecting RNA-seq (Supplementary Fig. 1b), ChIP–seq (Supplementary Fig. 1d–f), CAGE-seq (Supplementary Fig. 2) and ATAC-seq (Supplementary Fig. 1c) data. The DCC continues to be periodically updated (Extended Data Fig. 1e) and is openly accessible to the community for downloading data and uploading new datasets (Supplementary Videos 1 and 2).

The resulting data and reprocessed multiomic datasets represent a comprehensive annotation of the zebrafish genome during normal embryonic development and are available as a public track hub in the UCSC browser and uploadable to the Ensembl genome browser. Figure 1c provides an example developmentally regulated locus covering selected stages visualized by the Washington University Epigenome browser[29]. The tracks further include annotation of approximately 140,000 predicted ATAC-seq-supported developmental regulatory elements (PADRE) annotated by ChromHMM categories. The bulk sample-based predictions for regulatory elements are complemented with annotations of cell-type specificity of candidate regulatory elements provided by single-cell ATAC-seq[30] (Supplementary Videos 3–5).

Transcript annotation and core promoter characterization. As genome-wide transcriptome analyses[3,31–33] fail to annotate 5′ untranslated regions (UTRs) precisely, we used DANIO-CODE expression data to improve current Ensembl models of developmentally active genes. We utilized 139 developmental RNA-seq samples to identify 31,458 genes comprising 55,596 transcripts (Fig. 2a and Supplementary Table 2), among them 167 novel transcripts of uncertain coding potential (TUCP) and 726 long noncoding (lnc) RNA genes not previously annotated by Ensembl and supported by CAGE signals (Extended Data Fig. 2 and Supplementary Table 3). We mapped 5′ transcription start sites (TSSs) from 34 CAGE samples in 16 developmental stages (Fig. 2a). We applied promoter-calling criteria to CAGE data (Methods and Supplementary Fig. 2a–c), resulting in 22,500 active promoters per CAGE sample on average, corresponding to 16,303 genes (Supplementary Table 4), and adding 4,070 novel promoters to 18,461 previously annotated Ensembl TSSs (GRCz10). To supplement the promoterome with *cis*-regulatory sites, we curated 581 regulatory motifs representing 814 zebrafish transcription factor (TFs), and predicted binding sites for these motifs across all promoters (Methods).

Our above definition of promoters at single-nucleotide resolution may offer important guidance for promoter-targeted gene manipulation. For instance, gene promoter targeting for transcription block may be useful in reverse genetic experiments to avoid mutant RNA-mediated genetic compensation, which may mask mutant phenotypes and hinder dissection of gene function[34]. We compared Ensembl's RNA-seq-based TSS with our CAGE-seq-based TSS and found a substantial discrepancy in position (Fig. 2b and Extended Data Fig. 3a), potentially impacting guide RNA design for CRISPR–Cas targeting. Multiple dCas guide positions were designed and their impact on expression reduction with increased distance between the guide target and dominant CAGE-defined TSS was tested. Efficiency of dCas inhibition was higher when CAGE dominant, compared to Ensembl, start sites were used (Fig. 2c,d and Supplementary Table 5), demonstrating the importance of accurate TSS detection and the improved accuracy of CAGE over the current Ensembl pipeline in promoter detection.

Using these data we identified 1,293 multipromoter genes (Supplementary Table 6), where 1,176 genes had one reference and one alternative promoter and 117 genes had two or more alternative promoters. Correlation of expression levels of reference and alternative promoter pairs indicated both convergent (cyan in Fig. 3a,b) and divergent (brown) dynamics during development. The expression of reference promoters was on average higher than those of alternative promoters (Extended Data Fig. 3b). Among 978 transcript pairs with full-length coding sequence annotation, 373 (38%) of the alternative promoters affected only the 5′ UTR (for example, *dag1*; Extended Data Fig. 3c), whereas the remaining 605 altered the N-terminal protein sequence (for example, *bmp6*; Extended Data Fig. 3d). We analyzed mouse CAGE-seq[28] data from comparable embryonic stages and annotated 1,779 multipromoter genes (Extended Data Fig. 3e and Supplementary Table 7). About one-third (294; 30%) of identified mouse orthologs of zebrafish multipromoter genes (974; 75%) utilized alternative promoters. Orthologs of multipromoter genes were likely ($P = 2.7 \times 10^{-5}$; Fisher's exact test) to be expressed in similar stages and highly likely ($P = 3.24 \times 10^{-58}$; Fisher's exact test) to have multiple promoters in mouse. Multipromoter genes were enriched in KEGG signaling pathways in zebrafish (Fig. 3c) and mouse (Supplementary Table 8), suggesting vertebrate conservation of alternative promoters in signal transduction-associated genes.

Precision promoter annotation and expression dynamics allow exploitation of this resource to predict TF activity regulating the promoters. We implemented Motif Activity Response Analysis (MARA)[35,36] for zebrafish. MARA models promoter expression dynamics in terms of the annotated TF binding sites, to infer which TFs most substantially drive expression changes during development. Figure 3d shows the inferred activity profiles of three TFs with strong effects on genome-wide expression patterns. While Tead3 targets are upregulated from gastrulation onwards, Tgif1 targets are transiently downregulated and NF-Y targets are downregulated from the sphere stage onwards, consistent with the known

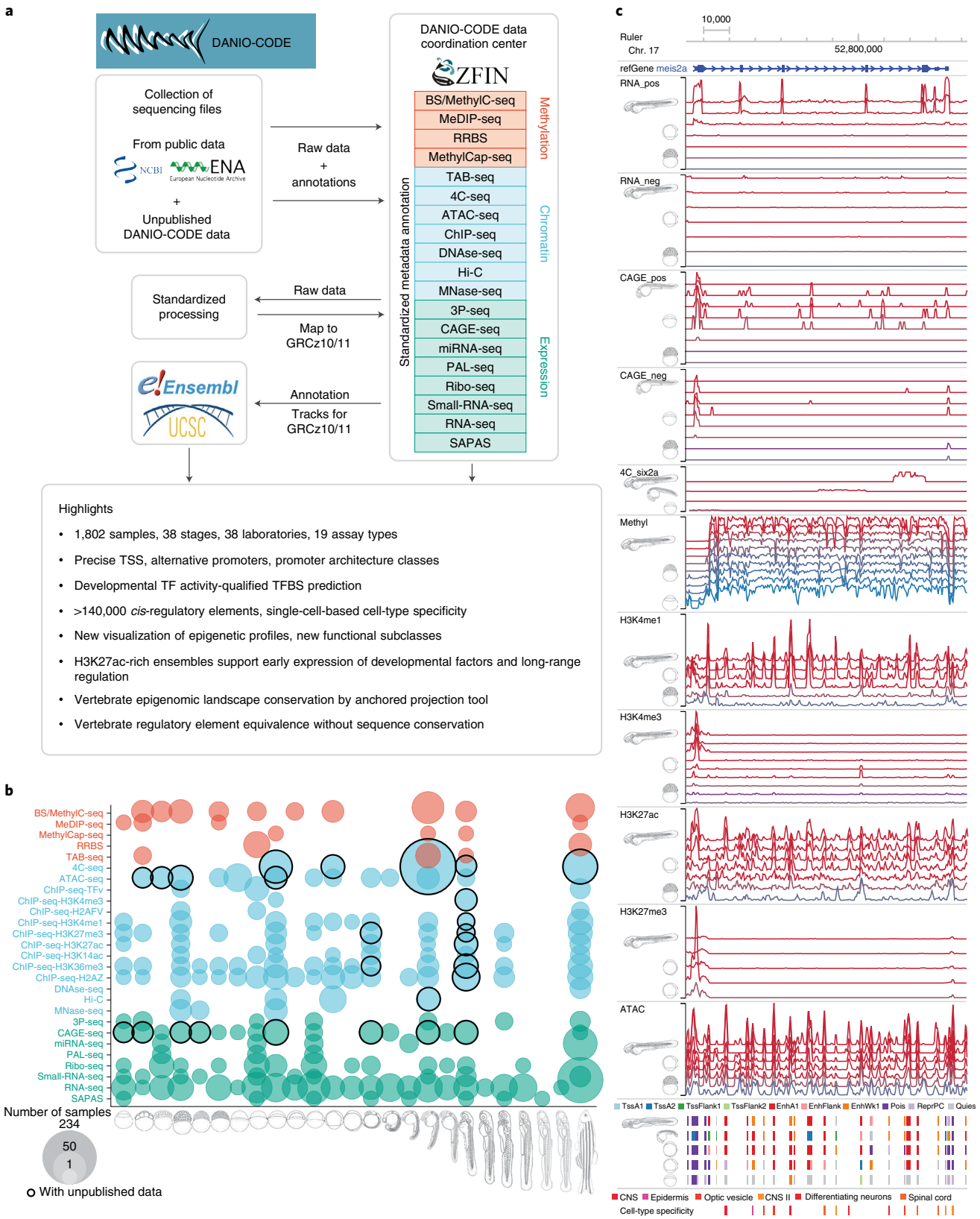


Fig. 1 | Comprehensive collection and annotation of zebrafish developmental genomic data. a, Collection and manual annotation processes of datasets with the DANIO-CODE DCC with highlights of key findings. **b**, Extent of the open repository for developmental multiomic data for zebrafish with assay type (*y* axis) and developmental stage (*x* axis). Data first reported in this study are highlighted with black circles. **c**, Visualization of temporal dynamics of selected transcriptomic and epigenomic features during development at a developmentally active locus. Coloring of tracks represents developmental series from maternal (blue) to zygotically active stages of embryogenesis (red). Symbols and track colors indicate representative stages (Extended Data Fig. 1d). CNS, central nervous system.

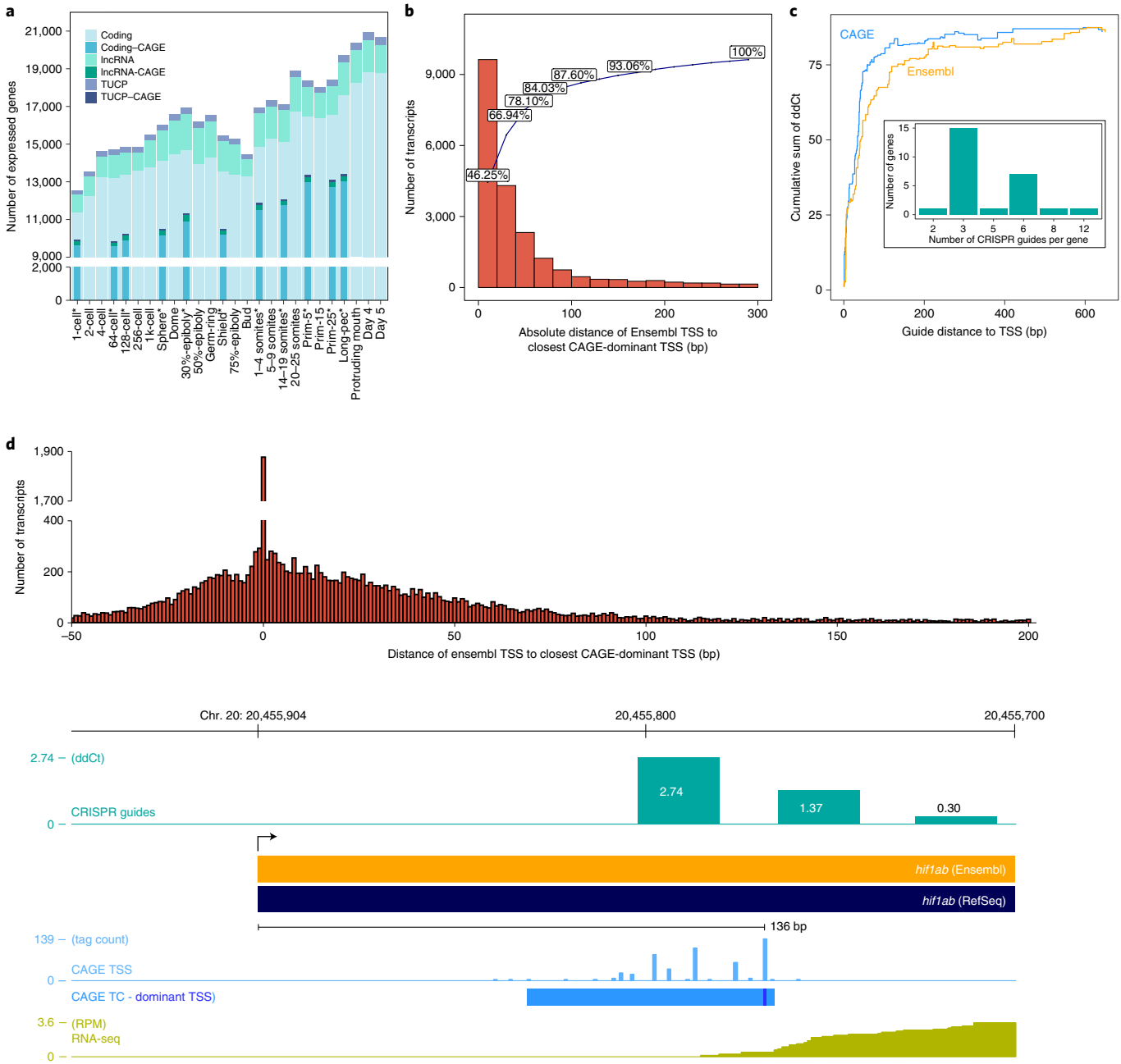


Fig. 2 | Transcript categories and single-nucleotide resolution 5′ end verification during development. a, DANIO-CODE transcript 5′ ends supported by CAGE TSS during stages of development. **b**, Distribution of absolute distance of Ensembl TSSs to CAGE-dominant TSSs in the Prim-5 stage. **c**, Relationship between guide distance to TSSs and ddCt. Inset: number of dCas guides for all 26 tested genes. **d**, CAGE-defined TSSs increase the accuracy of promoter identification and support dCas inhibition guide reagent designs. Distance between Ensembl TSSs and CAGE-dominant TSSs (top). Genome view with CRISPR guide position and efficacy, Ensembl and RefSeq transcripts, CAGE and RNA-seq expression (bottom).

activities of these TFs[37–41] (Extended Data Fig. 4 and Supplementary Table 9). MARA predicts substantially changing regulatory activities for regulatory motifs and assigns candidate regulator TFs to promoters (Fig. 3e). We have integrated our zebrafish annotations into the ISMARA webserver (ismara.unibas.ch) to allow this activity analysis on any RNA-seq data.

Classification of genomic regulatory regions in development. Next, we aimed to generate a comprehensive atlas of zebrafish developmental regulatory elements. We defined reproducible ATAC-seq[42] peaks as PADREs in four pre-zygotic genome activation (ZGA) and seven post-ZGA stages, which we further classified on the basis of the presence of four histone marks using ChromHMM[43,44] in five post-ZGA stages (Fig. 4a, Supplementary Fig. 3 and Extended Data Fig. 5a).

To examine the developmental dynamics of PADREs, we developed a UMAP-based method (Methods and Extended Data Fig. 6a–c) that can identify known functional classes and potentially novel subclasses during development. The UMAP plot of PADREs (Fig. 4b and Extended Data Fig. 6d) separated most ChromHMM-derived functional classes, including promoters from enhancers (Fig. 4c). Near-symmetry around the *y* axis reflects strand directionality and

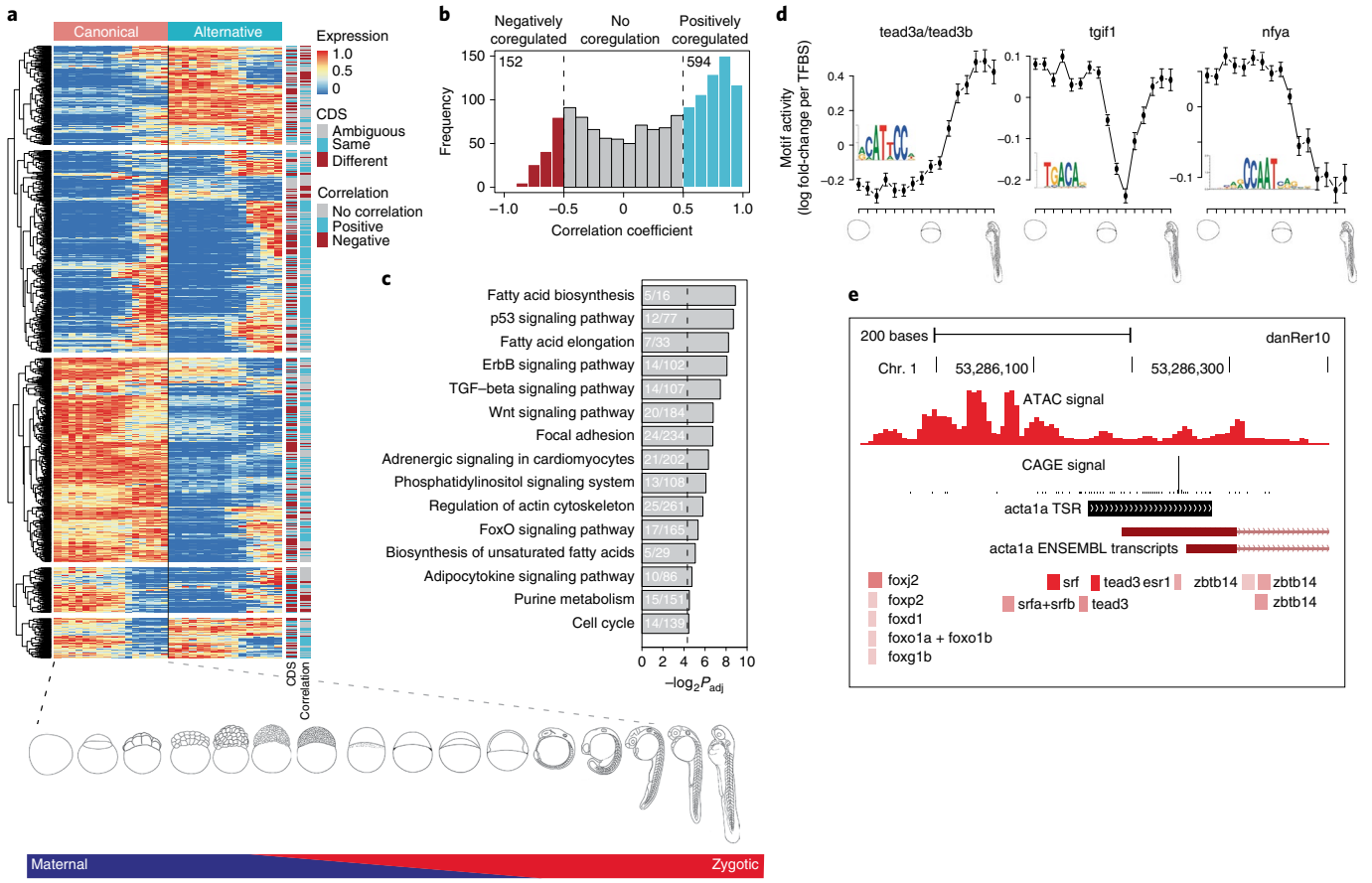


Fig. 3 | Alternative promoter usage and motif activity response analysis. a, Heat map shows the dynamics of expression levels of reference and alternative promoters across 16 developmental stages represented as images. Expression levels are scaled in the range of 0–1 for each row. Reference and alternative transcripts using the same and different coding sequence (CDS) starts are denoted. Transcript pairs without full CDS annotation are denoted as ambiguous. **b**, Distribution of correlation coefficient of expression levels of promoters across 16 developmental stages. **c**, Enrichment of KEGG pathways on multipromoter genes. The adjusted *P* value cut-off is 0.05, denoted by a vertical dashed line. The number of genes in KEGG pathways and those overlapping with multipromoter genes is shown inside the bars, **d**, MARA motif activity plots of three TF motifs across development. Posterior means and standard deviations (depicted as error bars) are based on analysis of the expression levels of all $n = 27{,}781$ promoters for each sample. Motif logos are depicted as insets. **e**, Genome browser view of the *actin alpha 1a* promoter. From the top: ATAC signal, CAGE signal, a single TSR (black), two Ensembl transcripts (dark red) and TFBSs predicted to regulate this TSR (red) are shown. Color intensities of the TFBSs reflect MARA scores of predicted regulatory role of TFs.

was most prominent among promoters (Fig. 4d). Two prominent clusters, which stretched upward and downward from the right apex and bear no chromatin marks, are enriched for the CTCF motif with well-positioned flanking nucleosomes[45] (Fig. 4e and Supplementary Fig. 4). Enhancer predictions were validated with two independent sets: (1) enhancers with bidirectional enhancer RNA signals[46] called from nuclear CAGE; and (2) a manually curated catalog of published enhancers functionally validated in transgenic reporter assays (Supplementary Table 10). Both colocalized with enhancer-classified PADREs on the UMAP (Fig. 4f,g and Extended Data Fig. 5c,d), demonstrating the utility of the method. DNA methylation analysis revealed CG-rich, promoter-associated PADREs persistently hypomethylated across stages, and less CG-dense enhancer-associated PADREs gradually hypermethylated during development before becoming hypomethylated in adult somatic tissue. Dynamically methylated PADREs varied in the onset and degree of hyper/hypomethylation: for example, conserved phylotypic enhancers[11] commenced hypomethylation at the Prim-5 stage (Extended Data Fig. 5f).

Next, we assessed the evolutionary conservation of PADREs by overlapping with human conserved noncoding elements (CNEs) and calculating the phastCons score for each PADRE (Fig. 4h, top, and Extended Data Fig. 5b). Early-acting enhancers appear less conserved than those activated later (Fig. 4h, bottom left, and Extended Data Fig. 5e). phastCons scores of enhancers were higher on average than those of promoters (Fig. 4h, bottom right). Poised elements were the most conserved, suggesting that Polycomb-bound enhancers are a specific class critical for differentiation and organogenesis[17,47], and contributing to the hourglass model of development[48].

To assign cell-type specificity to PADREs we integrated them with Prim-5 single-cell ATAC-seq[30] data (Extended Data Fig. 7a). The majority of anatomical annotation overlapped with transgenically confirmed enhancers and PADRE functional annotation (Extended Data Fig. 7b and Supplementary Table 11). UMAP (Fig. 4i, right) revealed remarkable differences between cell types, both within the same tissue and across tissues. PADREs active in neural precursors of the developing central nervous system showed a threefold increase of H3K27ac compared to those active in differentiating neurons, confirming previous observations about heterogeneity of cell-type population and chromatin dynamics in the developing central nervous system[49,50]. In contrast, PADREs active in muscle cells carried levels of H3K27ac and H3K4me1 comparable to neural precursors, but distinct accessibility profiles (Fig. 4i, bottom).

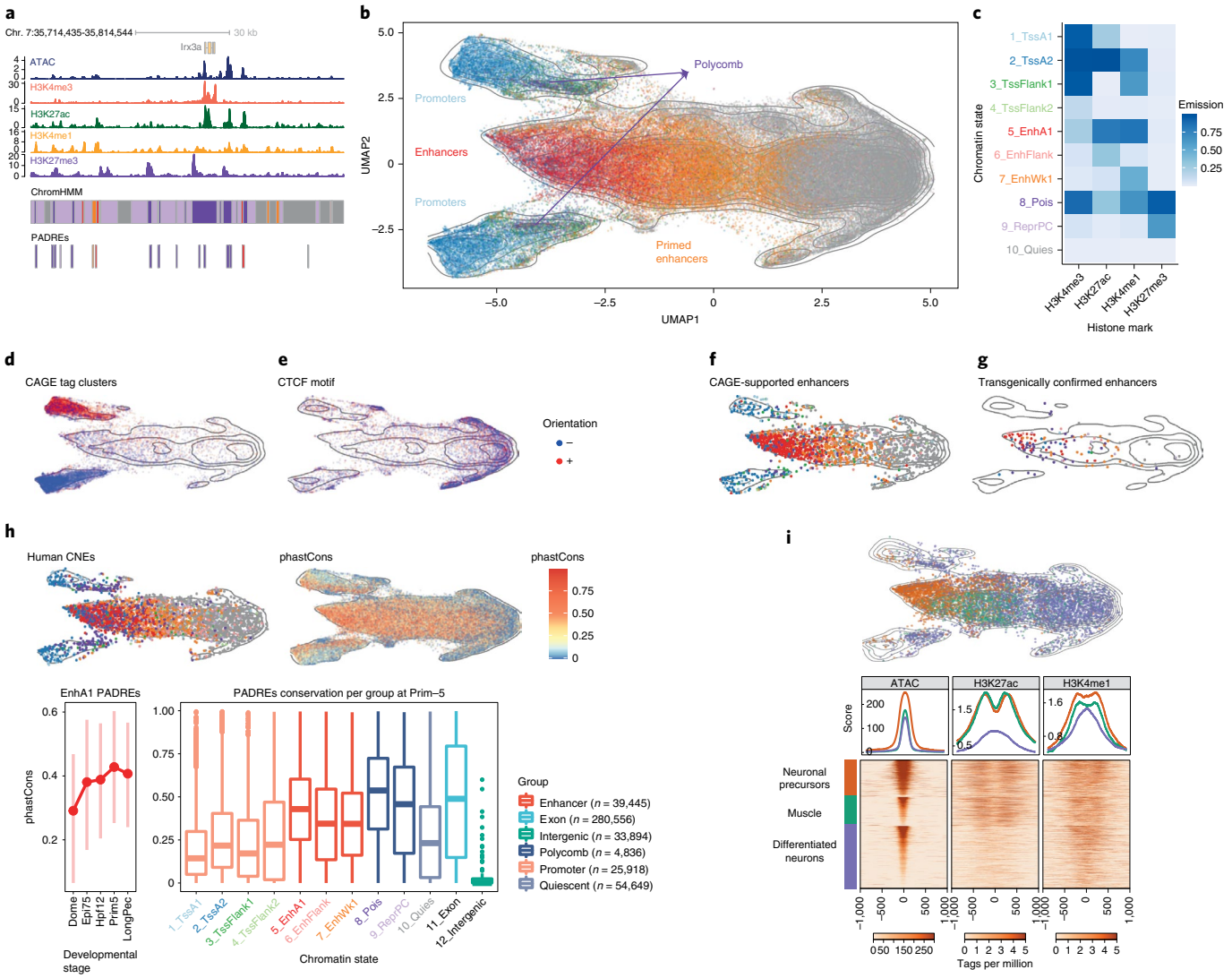


Fig. 4 | Classification of developmental *cis*-regulatory elements. a, Genome browser screenshot showing ChromHMM classification of PADREs, and respective histone post-translational modification signals used to define them. **b**, UMAP plot of PADREs at the Prim-5 stage. Each point represents one open chromatin region, colored by functional assignment. **c**, Occurrence probabilities of chromatin marks for ChromHMM states. The states function was manually assigned using The Roadmap Epigenomic annotations as reference. 1_TssA1, 2_TssA2: active TSS; 3_TssFlank1, 4_TssFlank2, TSS flanking region; 5_EnhA1, active enhancer; 6_EnhFlank, enhancer flanking region; 7_EnhWk1, primed enhancer; 8_Pois, poised elements; 9_PcRep, Polycomb-repressed regions; 10_Quies, quiescent state. **d**–**g**, UMAP plot showing PADREs overlapping with CAGE promoters (**d**), CTCF motif (**e**), eRNA enhancers (**f**) and transgenically validated enhancers (**g**). The transgenically validated enhancers are predominantly associated with enhancer-associated chromatin states (Supplementary Table 11). **h**, UMAP plot showing the mean phastCons score for each PADRE (top right) and overlap with human CNEs (top left). The bottom subpanel shows the distribution of phastCons scores of active enhancers throughout development (left, bars represents interquartile range), as well as the distribution of the phastCons score for PADREs separated by function at the Prim-5 stage. Two-sided Wilcoxon rank sum test was used to calculate *P* values between promoters and enhancers ($P = 2.2 \times 10^{-16}$) and enhancers and Polycomb-associated elements ($P = 2.2 \times 10^{-16}$). Exons and intergenic regions were added as reference (right) **i**, Position of cell-type-specific elements on the UMAP plot (top). ATAC, H3K27ac and H3K4me1 signals around the peak summit of cell-type-specific PADREs (bottom).

To understand the temporal dynamics of PADREs, we created a set of consensus PADREs (cPADREs), containing ~140,000 regions open in at least two neighboring stages (Supplementary Fig. 3a). We clustered nonpromoter cPADREs by chromatin accessibility into self-organizing maps (SOMs) (Extended Data Fig. 7c). Figure 5a (top) shows UMAP locations of 3 out of 16 SOM clusters, which demonstrate remarkable developmental chromatin changes, containing cPADREs active early and subsequently decommissioned (class 4), active from ZGA onwards (class 6) and late elements (class 14). Their chromatin profiles around ATAC-seq peaks were different, with only the early elements depleted of H3K27ac at their peak (Fig. 5a, bottom). With distinct chromatin and conservation profiles, early and late elements represent two separate classes of enhancers.

Finally, we explored the dynamics of PADREs without observable chromatin marks at any stage of development. 2,109 such regions were constitutively open throughout development (Supplementary Fig. 5a), which we termed constitutive orphan predicted elements (COPEs). They colocalized with constitutive SOM class 6 and 40% of them contained a CTCF motif (Figs. 5b, top, and 4e). In contrast, another nonmarked open chromatin set (11,044; termed dynamic orphan predicted elements; DOPES) was open only in

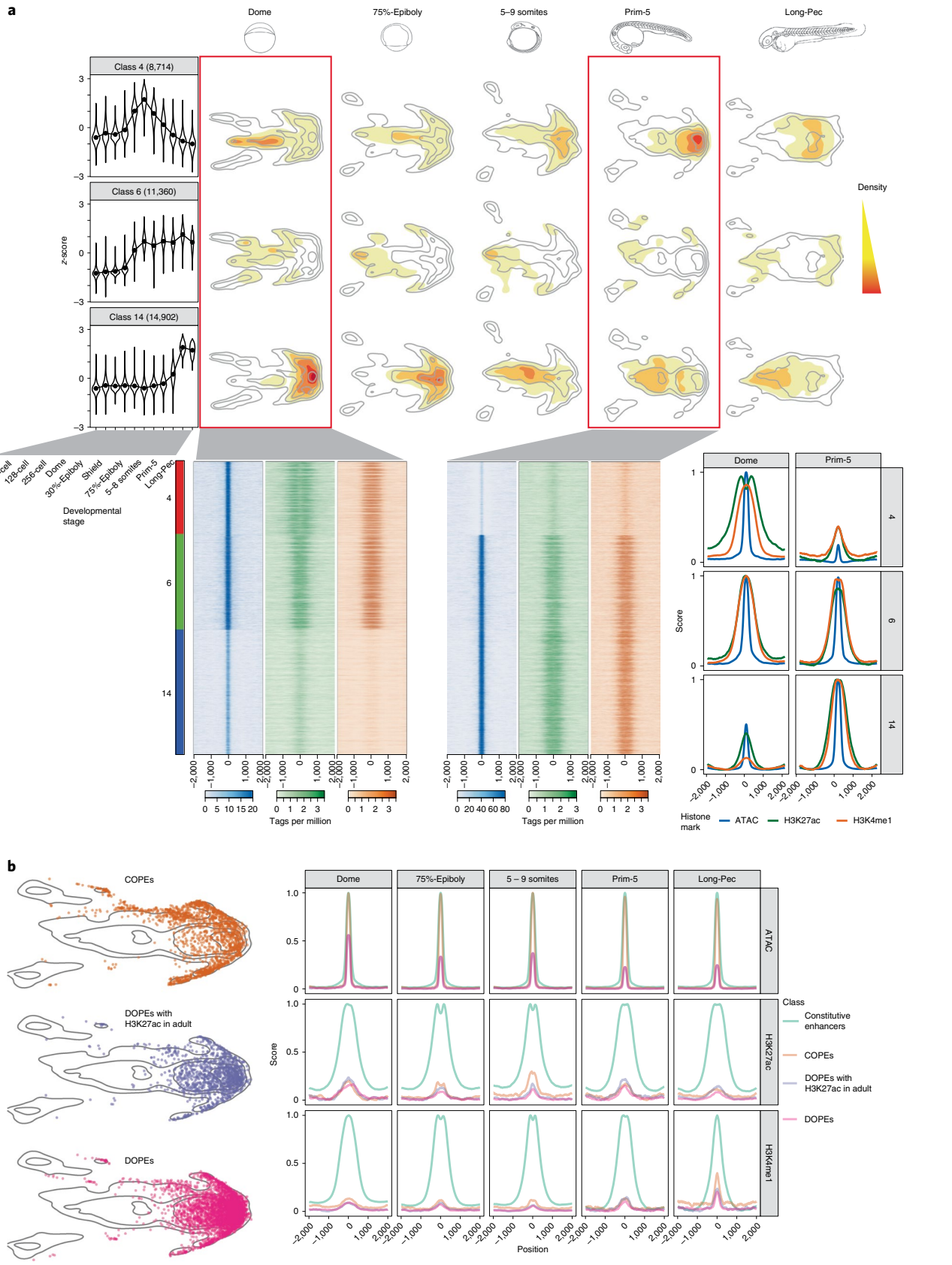


Fig. 5 | Developmental dynamics of PADREs. a, Openness profile of selected SOM classes (4: early; 6: post-ZGA constitutive; 14: late class), and their position density on the UMAP plots of different developmental stages (top). Heat map of signal intensity of ATAC, H3K27ac and H3K4me1 at the Dome and the Prim-5 stages, along with their respective profiles (bottom). **b**, Position of COPEs, DOPEs and DOPEs marked with H3K27ac in adult tissues on the UMAP plot (left). Profiles of ATAC, H3K27ac and H3K4me1 of COPEs, DOPEs, DOPEs marked in adult tissues, and other constitutive elements throughout development (right).

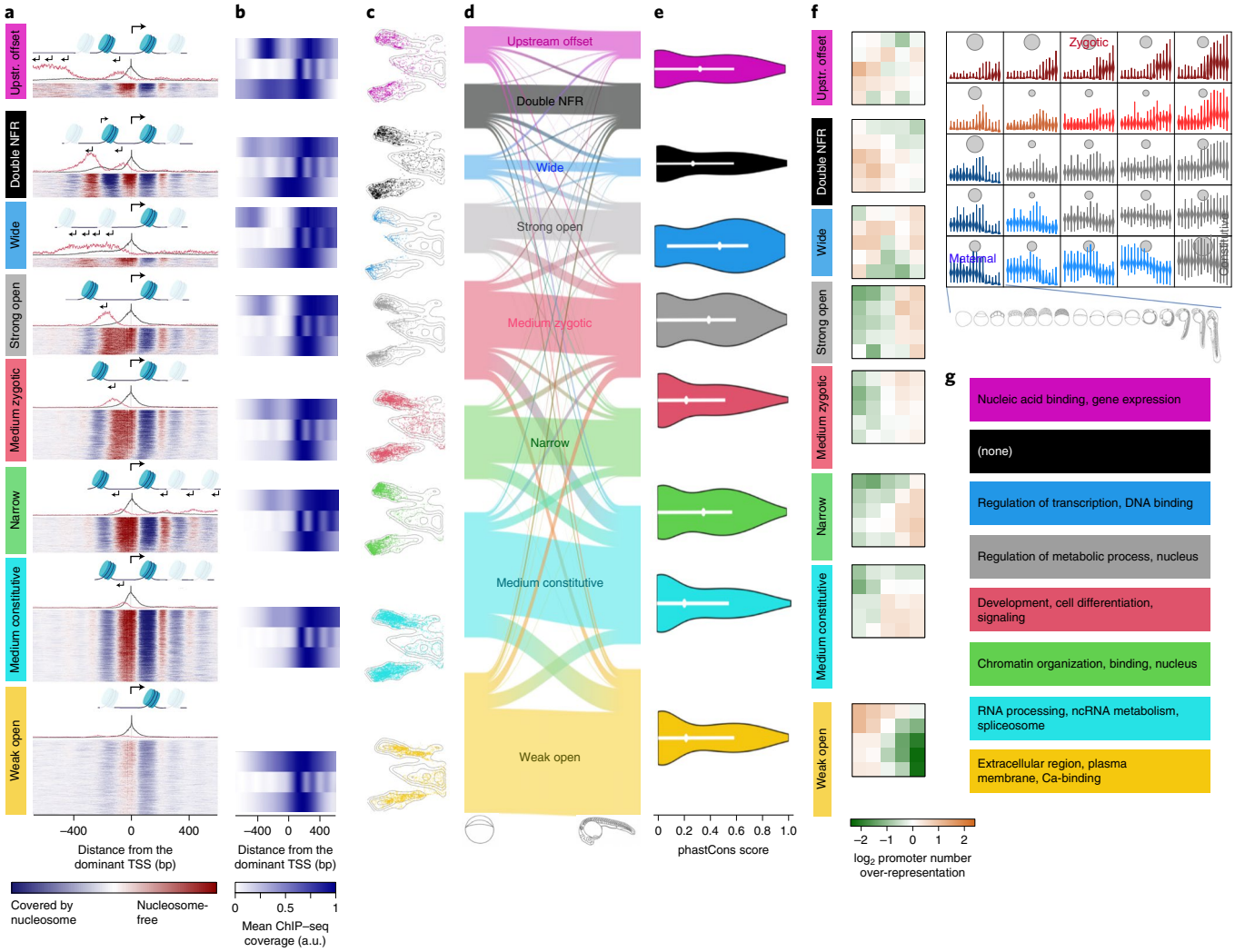


Fig. 6 | Chromatin architecture classification and developmental specialization of Pol II gene promoters. a, Heat map of chromatin accessibility profiles aligned to dominant TSS per promoter at the Prim-5 stage. Nucleosome-free regions (red) are superimposed with nucleosome positioning (blue). Stack height reflects number of promoters. Above each heat map, combined histograms of CAGE expression are shown. Black, forward TSSs; red, reverse orientation TSSs (the scale is amplified ×10 in relation to forward transcription). Nucleosome positioning is symbolized above alignments and black arrows indicate transcription direction; size indicates relative strength. Promoter configuration classes are color-coded consistently in all panels (including Supplementary Fig. 6) **b**, Aggregated H3K4me1, H3K4me3 (MNase-digested), H3K27ac ChIP-seq signals for classes as in **a** are aligned to dominant TSS. **c**, UMAP profiles of promoter classes at the Prim-5 stage. UMAPs are cropped to highlight promoter PADREs. **d**, Flow diagram indicates the relationship between promoter configuration class at the Dome stage (left edge, Supplementary Fig. 6) and the Prim-5 stage (right edge). Band width represents the number of promoters. **e**, Violin plot of phastCons vertebrate conservation distribution of promoters. Each class is aligned to **a**. **f**, Classification of promoter expression during development with SOMs. On the top right, 5×5 diagrams contain violin plots with stage-by-stage expression levels. Blue to red spectrum indicates maternal to zygotic expression dynamics of promoter clusters. Surface areas of gray circles indicate the number of promoters per cluster. Stages of development are symbolized below the SOM array. On the left, mustard: positive and green: negative color spectrum in SOMs indicates the enrichment in promoter overlap between promoter expression classes (SOMs) in each chromatin architecture class **a**. **g**, Enriched GO categories for each promoter architecture class.

specific developmental stages (Fig. 5b and Supplementary Fig. 5b). They were depleted of promoters, with only 65 (0.6%) overlapping CAGE promoters (Supplementary Table 12). Using data from ref. [24], we found that 2,513 DOPEs contained active chromatin marks later in adult tissues, but were open to the same extent as active enhancers already in the embryo (Supplementary Fig. 5c). As we are unaware of epigenetically 'orphaned' accessible elements in the development, whose chromatin opening precedes or is uncoupled from enhancer-associated histone mark deposition, this may represent a discovery of a previously unknown subtype of primed enhancers.

Developmental specialization of polymerase II gene promoters. To reveal any developmental promoter regulatory principles, we exploited the PADRE chromatin features to functionally classify CAGE-seq-defined active RNA polymerase II (Pol II) promoters. First, we characterized these promoters at Dome and Prim-5 stages on the basis of their chromatin accessibility at nucleosome resolution, revealing eight clusters (Fig. 6a, Supplementary Fig. 6a and Supplementary Table 13). We detected similar clusters in human embryonic stem cells (Supplementary Fig. 6b) indicating conservation of promoter chromatin architecture classes. The classes differed mostly in their upstream configuration, including the width of the

nucleosome-free region (NFR), the signal strength of the central NFR and the presence of upstream open regions (Fig. 6a), which followed GC content (Supplementary Fig. 6c). The NFRs only differed in their amplitude between 'medium constitutive' and 'weak open' (Supplementary Fig. 6a), with the latter either reflecting reduced promoter activity or promoters active only in a subset of cells. The NFR variations were characterized by histone marker presence and patterns of upstream opposite strand transcription (for example, 'upstream offset') with distinct distances between the main TSS and flanking nucleosomes (for example, 'wide' and 'strong open') and TSS profiles (Supplementary Fig. 6d). These classes showed notable differences in histone modification patterns (Fig. 6b), confirmed by the differing UMAP positions of promoter PADREs (Fig. 6c). Apart from 'weak open', each class produces antisense transcription (PROMPTs)[51–53], including 'double NFR', 'wide' and 'upstream offset' classes, which showed CAGE expression from both the main NFR and another upstream region, with sense transcription being stronger than antisense (Fig. 6a). Notably, the architecture classes remain stable over developmental time (Fig. 6d and Supplementary Fig. 6e), suggesting they represent distinct regulation mechanisms acting on the genes rather than stage-dependent promoter activity states. 'Wide' and 'strong open' classes contained the most conserved promoters (Fig. 6e and Supplementary Fig. 6f), and were enriched in transcription regulator genes (Fig. 6g and Supplementary Fig. 6g). However, promoter classes showed distinct dynamic temporal expression (Fig. 6f) with notable enrichment of the 'double NFR' class for maternally expressed genes, in contrast to the predominantly early and late zygotic 'weak open' and 'medium zygotic' classes, respectively. The promoter classes also showed distinct gene ontology (GO) enrichment categories (Fig. 6g). Overall, our approach offers a promoter architecture classification for zebrafish and indicates functional specialization and vertebrate conservation of promoter classes.

Developmental dynamics and locus organization of enhancers. Key genes regulating development are controlled by numerous long-range enhancers, which often overlap with highly conserved noncoding elements (HCNEs) within genomic regulatory blocks (GRBs)[15], which also often contain other 'bystander' genes that do not respond to those enhancers. The extent of GRBs coincides with those of topologically associating domains (TADs) around developmental genes[54] (Fig. 7a). We exploited DANIO-CODE annotations to characterize chromatin opening and interaction topology in those poorly understood loci, and their regulatory role in TADs.

We distinguished GRB TADs, characterized by a high density of extreme noncoding conservation, from non-GRB TADs. In the regions corresponding to late (Long-pec) embryo TADs, chromatin started opening at the boundaries as early as the Dome stage and remains open thereafter (Fig. 7b and Extended Data Fig. 8a). GRB TADs showed a strong increase in accessibility across the entire TAD, whereas in non-GRB TADs the increase was mild and occurred later (Fig. 7b). TADs started to form early but formed fully only at later developmental stages[55,56] (Extended Data Fig. 8b). We found more promoter-proximal enhancers in early stages and more distal enhancers in late stages, (Extended Data Fig. 8c), in line with similar findings by contact analysis[55].

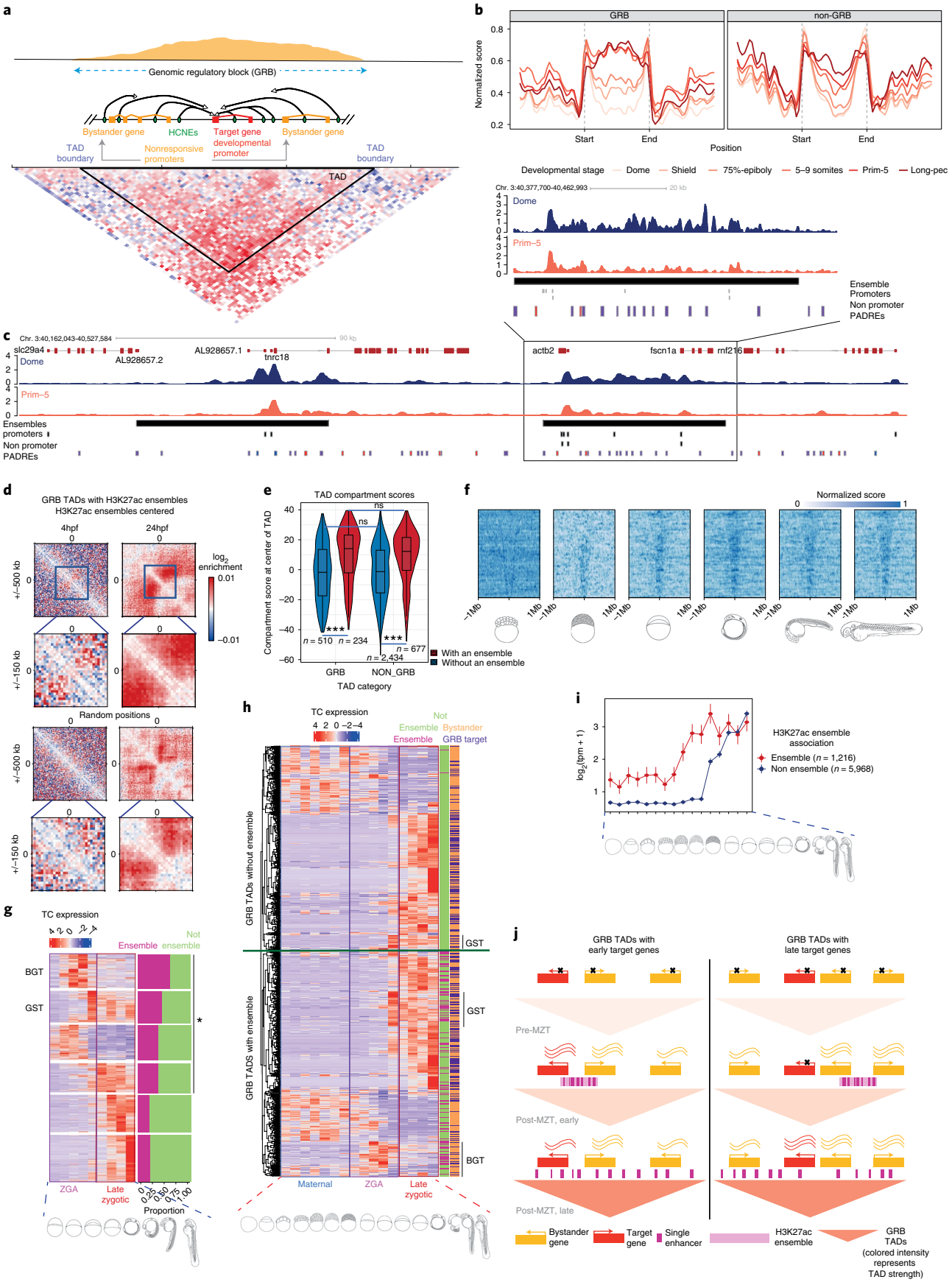
When we estimated the activity of enhancer candidates by H3K27ac in TADs, we observed that such elements in late stages are numerous, short and distributed throughout the entire TAD length. In contrast, many fewer PADREs were active early at Dome stage, and they often occurred in clusters with uninterrupted H3K27ac signal connecting them (Extended Data Fig. 8d,e and Fig. 7c). We detected ~1,600 such clusters[57], of which ~1,300 fell in TADs and were enriched in GRB TADs (Extended Data Fig. 8f). These clusters were reminiscent of super-enhancers[57,58], although more numerous than 231 reported in mouse[57] and 411 in zebrafish[56]. Given their unusual scale and early appearance before lineage determination (when previously reported super-enhancers appear), we distinguished them from super-enhancers and called them 'H3K27ac ensembles'. We hypothesized that they might be associated with the lack of fully formed TADs in the early stages, when enhancers are used proximally to early active promoters. To test this, we investigated the relationship between the chromatin interactions and activity of H3K27ac ensemble-associated genes during early versus late embryogenesis.

We found that promoters were enriched at the boundaries of H3K27ac ensembles (Extended Data Fig. 8g) and that the ensembles contain most candidate enhancer PADREs detected in early stages (Extended Data Fig. 8h). In contrast, the PADREs active only later in development represented long-range enhancers, distributed across the entire TAD (Extended Data Fig. 8d), and not enriched in ensembles (Extended Data Fig. 8h). Moreover, H3K27ac present along the entire length of the ensemble became restricted to individual peaks associated with PADREs by Prim-5 (Fig. 7c, zoomed-in panel).

Consistent with an H3K27ac ensemble role in early gene regulation, we observed increased Hi-C contacts within them at Dome in both GRB and non-GRB TADs. By Prim-5, strong contacts spread throughout the entire TAD (Fig. 7d and Extended Data Fig. 9a). TADs with H3K27ac ensembles present at Dome belonged to the active A compartment at Prim-5 (Fig. 7e), arguing for a role for H3K27ac ensembles in the timely opening of chromatin in their host TADs. Indeed, in GRB TADs, the H3K27ac mark propagated from H3K27ac ensembles to fill the entire TAD in later stages (Fig. 7f).

To examine how H3K27ac ensembles influence gene expression, we classified promoters within TADs by expression dynamics using SOM (Extended Data Fig. 9b). H3K27ac ensemble-associated promoters mostly sequestered into clusters, with the highest expression

Fig. 7 | Dynamics and function of open chromatin and H3K27ac topology organization on early embryo development. a, Schematic representation of GRBs. Basic components of a GRB. GRB enhancers (green) regulating the target genes span the entire length of the GRB (middle). Typical density pattern of conserved noncoding elements in a GRB, most of which overlap enhancers (top). Hi-C contact matrix within a GRB (bottom). **b**, Chromatin opening profiles through developmental stages along TADs. **c**, Genome browser view of a GRB TAD showing H3K27ac signals in the Dome and the Prim-5 stages, H3K27ac ensembles (black bar), CAGE promoters (black blocs) and nonpromoter PADREs (blue active in the Dome stage, red active in the Prim-5 stage and purple PADREs active in both stages). A zoomed-in genome browser view of an H3K27ac ensemble (top, left). **d**, Aggregate contact enrichment centered on ensembles at stages as indicated. **e**, TAD compartment score distribution. Positive scores represent A compartments, while negative ones represent B compartments. The comparison was done using two-sided two-sample unpaired Wilcoxon test. **g**, Heat maps of H3K27ac signal across GRB TADs containing ensembles through developmental stages. TADs are ordered by their width in descending order and fixed on the TAD center. **h**, CAGE expression patterns of selected gene classes separated by SOM, with the highest and lowest ratios in ensemble-associated genes. Bar plot on the right shows the proportion of ensemble-associated genes in each class. BGT and GST classes are marked on the heat map **i**, Gene expression pattern of GRB target and bystander genes. The left side bar shows an ensemble association for each gene. The right side bar shows the target or bystander assignment for each gene. Genes in TADs with and without ensembles are separated by a green line. BST and GST classes are indicated on the side. **j**, Graph showing mean expression and standard error of GRB target genes associated and not associated with early H3K27ac ensembles. **k**, A model describing the influence of an H3K27ac ensemble on expression of GRB target genes. If the H3K27ac ensemble is in contact with the target gene, it can be expressed early on.


a
Genomic regulatory block (GRB)
Bystander gene
HCNEs
Target gene
developmental promoter
Bystander gene
TAD boundary
Nonresponsive promoters
TAD boundary
TAD
b
GRB
non-GRB
Normalized score
Start
End
Position
Developmental stage
Dome
Shield
75%-epiboly
5–9 somites
Prim-5
Long-pec
Chr. 3:40,377,700-40,462,993
20 kb
Dome
Prim–5
Ensemble Promoters
Non promoter PADREs
c
Chr. 3:40,162,043-40,527,584
90 kb
slc29a4
AL928657.2
AL928657.1
tnrc18
actb2
fscn1a
rnf216
Dome
Prim–5
Ensembles promoters
Non promoter PADREs
d
GRB TADs with H3K27ac ensembles
H3K27ac ensembles centered
4hpf
24hpf
+/–500 kb
+/–150 kb
log2 enrichment
0.01
–0.01
Random positions
e
TAD compartment scores
Compartment score at center of TAD
ns
***
n = 510
n = 234
n = 677
n = 2,434
GRB
NON_GRB
TAD category
With an ensemble
Without an ensemble
f
Normalized score
–1Mb
1Mb
g
TC expression
Not Ensemble ensemble
BGT
GST
ZGA
Late zygotic
Proportion
h
TC expression
Not Ensemble Bystander Ensemble GRB target
GRB TADs without ensemble
GRB TADS with ensemble
GST
GST
BGT
Maternal
ZGA
Late zygotic
i
log2(tpm + 1)
H3K27ac ensemble association
Ensemble (n = 1,216)
Non ensemble (n = 5,968)
j
GRB TADs with early target genes
GRB TADs with late target genes
Pre-MZT
Post-MZT, early
Post-MZT, late
Bystander gene
Target gene
Single enhancer
H3K27ac ensemble
GRB TADs (colored intensity represents TAD strength)

in early post-ZGA stages (Fig. 7g). We termed the top two H3K27ac ensemble-associated classes as blastula-gastrula transition (BGT) and gastrula-segmentation transition (GST) on the basis of the peak expression time. The two major gene classes in GRB TADs were ubiquitously expressed (GRB bystanders) and late zygotic expressed (likely GRB target genes). However, in GRB TADs with an ensemble, we observed a BGT gene class, not present in GRB TADs without an ensemble, as well as more genes in the GST class. Both classes were enriched in ensemble-associated genes (Fig. 7h). Moreover, there was a clear trend of earlier expression in H3K27ac ensemble-associated GRB target genes, compared to other GRB target genes (Fig. 7i), suggesting that ensembles participated in the activation of early-acting developmental genes, including those later dependent on long-range regulation. Moreover, if the target gene is not in contact with the H3K27ac ensemble, it can only become expressed once long-range interactions are present (Fig. 7j).

Functional conservation of epigenetic subdomains. Next, we investigated whether our annotation of noncoding elements could be exploited to predict functionally conserved *cis*-regulatory elements (CREs) among vertebrates. Existing comparative methods rely on direct alignments between species of interest[59,60]. However, the large evolutionary distance between fish and mammals limits the power of comparison, due to loss of noncoding sequence similarity. We developed a method to predict functional conservation across large evolutionary distances and genomic scales independent of direct sequence alignment, exploiting the fact that functional elements often maintain collinear syntenic positions, while their spacing scales with genome size, particularly in GRB TADs[15,54,61,62]. We selected 13 high-quality bridging species reference genomes and using stepped pairwise sequence alignment (Extended Data Fig. 10 and Methods), which allowed us to map coordinates between genomes of varying sizes, identified reference points (multispecies anchors; Fig. 8a) between genomes and enabling identification of syntenic regions through interpolation of relative syntenic positions between anchor points.

We then compared zebrafish and mouse GRB TADs, which differ in size approximately twofold (Extended Data Fig. 10a). We defined GRB TADs as the 1,000 TADs with the highest CNE density, split them into 1-kilobase (kb) bins, and mapped the bin centers from zebrafish to mouse. Using our multispecies approach over direct alignment reduced distances from the bin centers to their closest anchor by a factor of 16 in zebrafish and 29 in mouse (Fig. 8b).

We asked whether this method could discover conserved epigenomic subdomains by comparing epigenomic feature distribution across genomes. We used H3K27me3 ChIP–seq data from phylotypic stages in zebrafish (Prim-5) and mouse (E10.5; Methods). H3K27me3 coordinates from zebrafish were projected onto the mouse genome, recovering mouse H3K27me3 features in the corresponding region. An example at the *irx3a* locus (Fig. 8c) shows H3K27me3 enrichment correlates between zebrafish and mouse, even in the absence of direct sequence conservation. On a genome-wide level, H3K27me3 enrichment is substantially more likely to be shared between zebrafish and mouse for both directly alignable and nonalignable genomic regions (Extended Data Fig. 10e), suggesting epigenomic subdomains and functional elements can be conserved in location and span. We see more GRB TADs showing regions of strong similarity in H3K27me3 extent, while others, such as TADs containing *her9* or *celf5a*, show more zebrafish- or mouse-specific signal enrichment, and still others show little enrichment (Fig. 8d,e).

We next looked at conservation of functional elements marked by open chromatin. We classified zebrafish ATAC-seq peaks in the GRB TADs as directly conserved (DC) if they fall in a region of direct sequence alignment with mouse (16,188 elements, 11.5 %), indirectly conserved (IC) if they do not directly align (6,137 elements, 4.4%) but were alignable through bridging species and nonconserved (NC) for all other peaks (for example, *irx3a* in Fig. 8f). Notably, DC and IC elements shared regulatory features with their matched counterparts in mouse, including DNase hypersensitivity and ChromHMM feature classification, compared to NC elements (Fig. 8g,h). DC and IC regions were also more likely to share TF binding site (TFBS) motifs compared to nonoverlapping, randomly sampled mouse DNase-seq peaks within and across TAD boundaries (*cis* and *trans* in Fig. 8i and Supplementary Table 14). These results suggest a similar level of functional conservation of DC and IC elements, even though IC elements lack direct alignability. Next, we tested whether the early developmental H3K27ac ensembles detected in zebrafish embryos (Fig. 7) are conserved in mouse using our anchoring-based approach. As shown in Fig. 8j,k, H3K27ac signal in mouse was substantially enriched in zebrafish ensembles, suggesting these ensembles are evolutionarily conserved epigenetic subdomains in vertebrates. Genes associated with these conserved ensembles are listed in Supplementary Table 15. Our comparative epigenomic approach has maximized the identification of putative functional elements and epigenetic subdomains conserved between zebrafish and mouse, and highlights the utility of the DANIO-CODE annotations for discovery of vertebrate-conserved mechanisms.

Discussion

Here we describe the establishment and provision of a zebrafish developmental genomics resource as a track hub and downloadable resource within a data coordination center, which is designed for expansion by continued incorporation of future zebrafish genomic data. Our track hub allows visualization of developmental noncoding functional annotations in common genome browsers.

We have annotated over 140,000 candidate developmental CREs, including enhancers and promoters. There is a need for the classification of CREs to reflect their distinct temporal and spatial dynamics and modes of functionality. Recognizing this, we improved the classifications of enhancers and promoters with novel subcategories using dimensionality reduction on chromatin accessibility and nucleosome-level histone modifications. CRE subclasses include DOPEs and COPEs, which may carry as-yet unmapped histone marks[63] and merit further investigation. We demonstrate distinct local chromatin architecture of CREs in developmental stages and in developing cell types. Moreover, we classified promoters into potentially novel chromatin architecture classes, which we also detected in mammals, and which are distinctly used by subsets of genes. We have integrated our CRE annotations with chromosome topology and explored the dynamic organization of CRE interactions during development. We identified large H3K27ac marked ensembles, which are distinct from previously described super-enhancers targeting lineage-determining genes. We suggest ensembles function in nuclear topology organization at local interaction hubs around early active loci during the initial formation of TADs.

The datasets used in this study are bulk whole-embryo samples, which can mask chromatin state dynamics of rare cell populations or varying cell cycle states. Nevertheless, we were able to identify distinct subclasses of candidate CREs, such as promoter classes and COPEs, by comparing intersections of independent chromatin features. In this effort, stable chromatin features served as references to compare the varying dynamics of overlapping chromatin states, mitigating bulk averaging artifacts. The expansion of single-cell genomic data by the zebrafish community will help in further stratifying *cis*-regulatory classes. Meanwhile users are encouraged to browse tracks for enhancer marks emanating from small cell numbers that may be masked by thresholding, and to integrate tissue- and cell-type specificity information. Such integration will generate further layers of functional annotations, including TF expression and binding sites[64] and help in identification of gene regulatory network components acting in lineage determination[65].

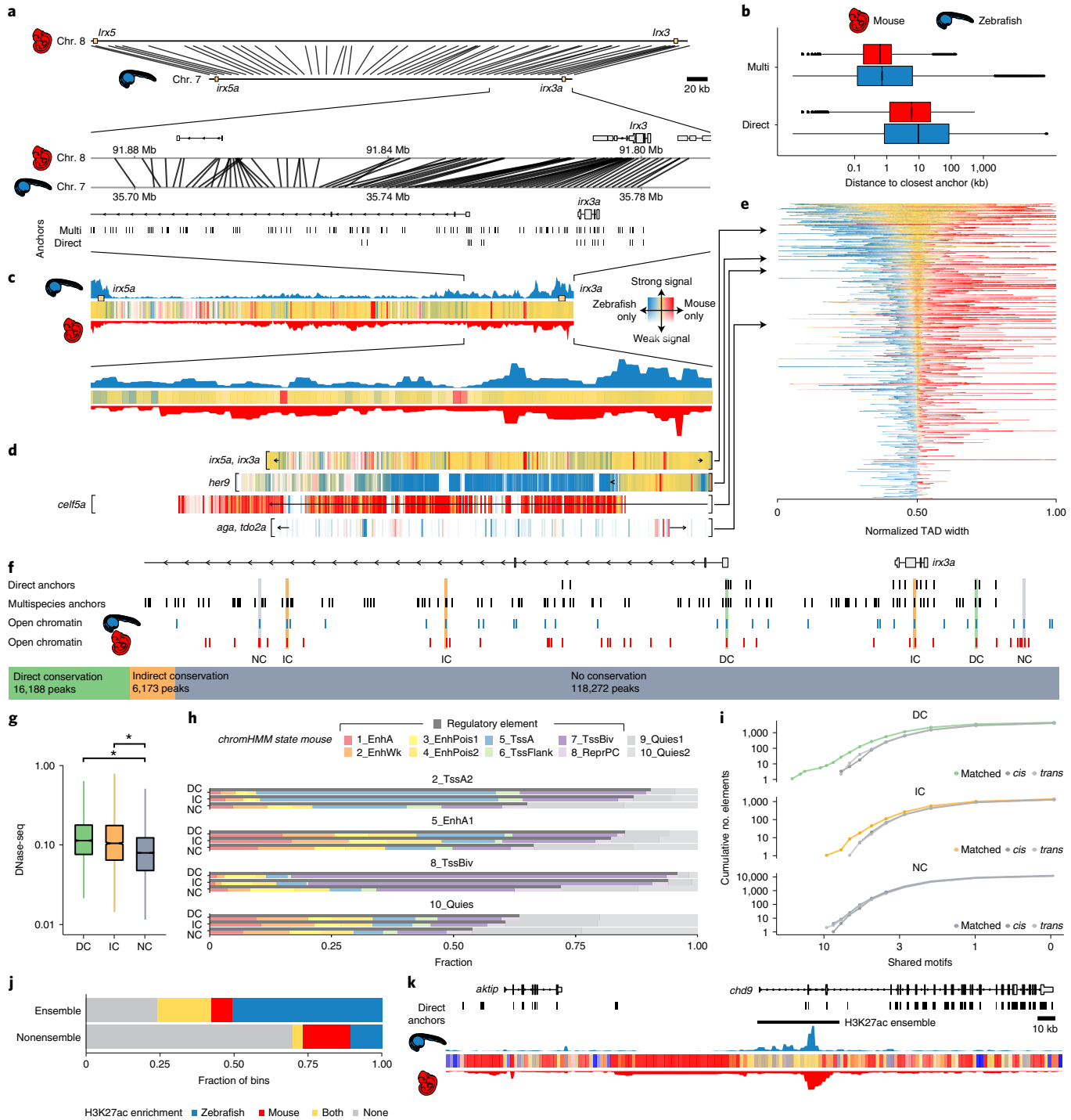


Fig. 8 | Synteny projections reveal conservation of epigenetic features between zebrafish and mouse. a, Cross-species comparison of the *irx3/5*(a) TAD between zebrafish and mouse and a zoom-in on the locus around *irx3*(a). Connecting lines represent projections of bin centers from zebrafish to mouse. **b**, Distribution of distances from the bin centers ($n = 528{,}830$) to their closest anchors in zebrafish (blue), and from their projections to their closest anchors in mouse (red), using the direct and the multispecies projection approach. **c**, Epigenetic comparison of the *irx3/5*(a) TAD. H3K27me3 overlap in mapped regions is indicated as colored bars (yellow, mutually enriched; blue, zebrafish specific; red, mouse specific; Methods). Opacity reflects signal amplitude and is proportional to the maximum H3K27me3 signal in both species. **d**, H3K27me3 overlap profiles for four selected GRB TADs. TAD boundaries are indicated with square brackets. **e**, H3K27me3 overlap profiles of all GRB TADs. TADs are ordered by their relative amount of shared signal. Bins are ordered by the amount of shared signal: bins with shared signal appear in the middle, bins with zebrafish- and mouse-specific signals are left and right, respectively. A view of the TADs with their genomic bin order is given in Extended Data Fig. 10d,f. Classification of zebrafish ATAC-seq peaks in the *irx3a* locus into DC, IC and NC on the basis of overlaps with direct anchors, multispecies anchors and mouse DNase-seq peak projections (Methods). **g**, Distribution of DNase-seq signal in the mouse genome around the projected regions of the zebrafish ATAC-seq peaks ($n = 140{,}633$). Asterisks above the bars indicate the effect size category based on Cohen's *d*: very small (not indicated), small (*), medium (**), large (***), very large (****). **h**, Cross-species comparison of ChromHMM functional states. **i**, Cumulative distribution of shared motifs in mouse DNase-seq peaks overlapping zebrafish ATAC-seq peaks. **j**, H3K27ac enrichment (signal ≥80th percentile) within ($n = 11{,}083$) and outside ($n = 93{,}020$) of enhancer ensembles ($P < 2.2 \times 10^{-16}$, Fisher's exact test). **k**, Cross-species comparison of H3K27ac profile around an H3K27ac ensemble neighboring the zebrafish *aktip* gene.

The DCC and the functional annotation track hub will thus serve as a foundation for future single-cell studies of transcriptomes[66–68], open chromatin[30] and others, as demonstrated with single-cell ATAC-seq data here. The functional annotations presented will also aid in targeted manipulation of genomic elements. For example, our high-resolution promoter annotation will aid reagent design for gene regulation assays[69], transgenic cell labeling[70] and transcription blocking.

The utility of our functional annotations extends well beyond zebrafish development. We developed an approach that detects functional equivalence of regulatory landscapes in the absence of sequence conservation. Our multispecies anchoring approach facilitated the identification of nonsequence conserved positional equivalents with enrichment for shared epigenetic domains (H3K27me3 and H3K27ac) and syntenic enhancer TFBS content, highlighting the predictive value and functional relevance of epigenetic subdomains within syntenic TADs. This zebrafish resource thus expands on and complements the existing functional genome mapping efforts in mammals and modENCODE species.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-022-01089-w.

Received: 9 August 2021; Accepted: 3 May 2022;
Published online: 04 July 2022

References

1. Patton, E. E. & Tobin, D. M. Spotlight on zebrafish: the next wave of translational research. *Dis. Models Mechanisms* **12**, dmm039370 (2019).
2. Howe, D. G. et al. The zebrafish model organism database: new support for human disease models, mutation details, gene expression phenotypes and searching. *Nucleic Acids Res.* **45**, D758–D768 (2017).
3. Howe, K. et al. The zebrafish reference genome sequence and its relationship to the human genome. *Nature* **496**, 498–503 (2013).
4. Bogdanovic, O. et al. Dynamics of enhancer chromatin signatures mark the transition from pluripotency to cell specification during embryogenesis. *Genome Res.* **22**, 2043–2053 (2012).
5. Murphy, P. J., Wu, S. F., James, C. R., Wike, C. L. & Cairns, B. R. Placeholder nucleosomes underlie germline-to-embryo DNA methylation reprogramming. *Cell* **172**, 993–1006.e13 (2018).
6. Vastenhouw, N. L. et al. Chromatin signature of embryonic pluripotency is established during genome activation. *Nature* **464**, 922–926 (2010).
7. Haberle, V. et al. Two independent transcription initiation codes overlap on vertebrate core promoters. *Nature* **507**, 381–385 (2014).
8. Bazzini, A. A., Lee, M. T. & Giraldez, A. J. Ribosome profiling shows that miR-430 reduces translation before causing mRNA decay in zebrafish. *Science* **336**, 233–237 (2012).
9. Nepal, C. et al. Dual-initiation promoters with intertwined canonical and TCT/TOP transcription start sites diversify transcript processing. *Nat. Commun.* **11**, 168 (2020).
10. Zhao, L., Wang, L., Chi, C., Lan, W. & Su, Y. The emerging roles of phosphatases in Hedgehog pathway. *Cell Commun. Signal.* **15**, 35 (2017).
11. Bogdanović, O. et al. Active DNA demethylation at enhancers during the vertebrate phylotypic period. *Nat. Genet.* **48**, 417–426 (2016).
12. Jiang, L. et al. Sperm, but not oocyte, DNA methylome is inherited by zebrafish early embryos. *Cell* **153**, 773–784 (2013).
13. Potok, M. E., Nix, D. A., Parnell, T. J. & Cairns, B. R. Reprogramming the maternal zebrafish genome after fertilization to match the paternal methylation pattern. *Cell* **153**, 759–772 (2013).
14. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
15. Kikuta, H. et al. Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res.* **17**, 545–555 (2007).
16. Gehrig, J. et al. Automated high-throughput mapping of promoter-enhancer interactions in zebrafish embryos. *Nat. Methods* **6**, 911–916 (2009).
17. Rada-Iglesias, A. et al. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279–283 (2010).
18. Spieler, D. et al. Restless legs syndrome-associated intronic common variant in Meis1 alters enhancer function in the developing telencephalon. *Genome Res.* **24**, 592–603 (2014).
19. Encode Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
20. Dixon, J. R. et al. Chromatin architecture reorganization during stem cell differentiation. *Nature* **518**, 331–336 (2015).
21. Kundaje, A. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
22. Gerstein, M. B. et al. Integrative analysis of the *Caenorhabditis elegans* genome by the modENCODE project. *Science* **330**, 1775–1787 (2010).
23. Roy, S. et al. Identification of functional elements and regulatory circuits by *Drosophila* modENCODE. *Science* **330**, 1787–1797 (2010).
24. Yang, H. et al. A map of *cis*-regulatory elements and 3D genome structures in zebrafish. *Nature* **588**, 337–343 (2020).
25. Tan, H., Onichtchouk, D. & Winata, C. DANIO-CODE: toward an encyclopedia of DNA elements in zebrafish. *Zebrafish* **13**, 54–60 (2016).
26. Hortenhuber, M., Mukarram, A. K., Stoiber, M. H., Brown, J. B. & Daub, C. O. *-DCC: A platform to collect, annotate, and explore a large variety of sequencing experiments. *GigaScience* **9**, giaa024 (2020).
27. Encode Project Consortium et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
28. The FANTOM Consortium and the RIKEN PMI and CIST. A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).
29. Li, D., Hsu, S., Purushotham, D., Sears, R. L. & Wang, T. WashU epigenome browser update 2019. *Nucleic Acids Res.* **47**, W158–W165 (2019).
30. McGarvey, A, C. et al. Single-cell-resolved dynamics of chromatin architecture delineate cell and regulatory states in zebrafish embryos. *Cell Genom.* **2**, 100083 (2022).
31. Pauli, A. et al. Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res.* **22**, 577–91 (2012).
32. White, R. J. et al. A high-resolution mRNA expression time course of embryonic development in zebrafish. *eLife* **6**, e30860 (2017).
33. Lawson, N. D. et al. An improved zebrafish transcriptome annotation for sensitive and comprehensive detection of cell type-specific genes. *eLife* **9**, e55792 (2020).
34. El-Brolosy, M. A. et al. Genetic compensation triggered by mutant mRNA degradation. *Nature* **568**, 193–197 (2019).
35. The FANTOM Consortium and Riken Omics Science Center The transcriptional network that controls growth arrest and differentiation in a human myeloid leukemia cell line. *Nat. Genet.* **41**, 553–562 (2009).
36. Balwierz, P. J. et al. ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs. *Genome Res.* **24**, 869–884 (2014).
37. Astone, M. et al. Zebrafish mutants and TEAD reporters reveal essential functions for Yap and Taz in posterior cardinal vein development. *Sci. Rep.* **8**, 10189 (2018).
38. Chae, H. D., Yun, J., Bang, Y. J. & Shin, D. Y. Cdk2-dependent phosphorylation of the NF-Y transcription factor is essential for the expression of the cell cycle-regulatory genes and cell cycle G1/S and G2/M transitions. *Oncogene* **23**, 4084–4088 (2004).
39. Hu, Q., Lu, J. F., Luo, R., Sen, S. & Maity, S. N. Inhibition of CBF/NF-Y mediated transcription activation arrests cells at G2/M phase and suppresses expression of genes activated at G2/M phase of the cell cycle. *Nucleic Acids Res.* **34**, 6272–6285 (2006).
40. Powers, S. E. et al. Tgif1 and Tgif2 regulate Nodal signaling and are required for gastrulation. *Development* **137**, 249–259 (2010).
41. Szklarczyk, D. et al. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* **49**, D605–D612 (2021).
42. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
43. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).
44. Ernst, J. & Kellis, M. Chromatin-state discovery and genome annotation with ChromHMM. *Nat. Protoc.* **12**, 2478–2492 (2017).
45. Fu, Y., Sinha, M., Peterson, C. L. & Weng, Z. The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genet.* **4**, e1000138 (2008).
46. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
47. Crispatzu, G. et al. The chromatin, topological and regulatory properties of pluripotency-associated poised enhancers are conserved in vivo. *Nat. Commun.* **12**, 4344 (2021).

48. Tena, J. J. et al. Comparative epigenomics in distantly related teleost species identifies conserved *cis*-regulatory nodes active during the vertebrate phylotypic period. *Genome Res.* **24**, 1075–1085 (2014).
49. Raj, B. et al. Emergence of neuronal diversity during vertebrate brain development. *Neuron* **108**, 1058–1074.e6 (2020).
50. Lister, R. et al. Global epigenomic reconfiguration during mammalian brain development. *Science* **341**, 1237905 (2013).
51. Core, L. J., Waterfall, J. J. & Lis, J. T. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science* **322**, 1845–1848 (2008).
52. Seila, A. C. et al. Divergent transcription from active promoters. *Science* **322**, 1849–1851 (2008).
53. Buratowski, S. Transcription. Gene expression–where to start? *Science* **322**, 1804–1805 (2008).
54. Harmston, N. et al. Topologically associating domains are ancient features that coincide with Metazoan clusters of extreme noncoding conservation. *Nat. Commun.* **8**, 441 (2017).
55. Kaaij, L. J. T., van der Weide, R. H., Ketting, R. F. & de Wit, E. Systemic loss and gain of chromatin architecture throughout zebrafish development. *Cell Rep.* **24**, 1–10.e4 (2018).
56. Wike, C. L. et al. Chromatin architecture transitions from zebrafish sperm through early embryogenesis. *Genome Res.* **31**, 981–994 (2021).
57. Whyte, W. A. et al. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307–319 (2013).
58. Hnisz, D. et al. Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947 (2013).
59. Villar, D. et al. Enhancer evolution across 20 mammalian species. *Cell* **160**, 554–566 (2015).
60. Xiao, S. et al. Comparative epigenomic annotation of regulatory DNA. *Cell* **149**, 1381–1392 (2012).
61. Crollius, H. R., Gilardi-Hebenstreit, P., Torbey, P. & Clément, Y. Enhancer-gene maps in the human and zebrafish genomes using evolutionary linkage conservation. *Nucleic Acids Res.* **48**, 2357–2371 (2020).
62. Engstrom, P. G., Ho Sui, S. J., Drivenes, O., Becker, T. S. & Lenhard, B. Genomic regulatory blocks underlie extensive microsynteny conservation in insects. *Genome Res.* **17**, 1898–1908 (2007).
63. Pradeepa, M. M. et al. Histone H3 globular domain acetylation identifies a new class of enhancers. *Nat. Genet.* **48**, 681–686 (2016).
64. Fornes, O. et al. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **48**, D87–D92 (2020).
65. Davidson, E. H. Emerging properties of animal gene regulatory networks. *Nature* **468**, 911–920 (2010).
66. Briggs, J. A. et al. The dynamics of gene expression in vertebrate embryogenesis at single-cell resolution. *Science* **360**, eaar5780 (2018).
67. Farnsworth, D. R., Saunders, L. M. & Miller, A. C. A single-cell transcriptome atlas for zebrafish development. *Dev. Biol.* **459**, 100–108 (2020).
68. Farrell, J. A. et al. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science* **360**, eaar3131 (2018).
69. Housden, B. E. et al. Loss-of-function genetic tools for animal models: cross-species and cross-platform differences. *Nat. Rev. Genet.* **18**, 24–40 (2016).
70. Sakaue-Sawano, A. et al. Visualizing spatiotemporal dynamics of multicellular cell-cycle progression. *Cell* **132**, 487–498 (2008).

[1]MRC London Institute of Medical Sciences, London, UK. [2]Institute of Clinical Sciences, Faculty of Medicine, Imperial College London, Hammersmith Hospital Campus, London, UK. [3]Department of Biosciences and Nutrition, Karolinska Institutet, NEO, Huddinge, Sweden. [4]Institute of Cancer and Genomic Sciences, Birmingham Centre for Genome Biology, College of Medical and Dental Sciences, University of Birmingham, Birmingham, UK. [5]Max Planck Institute for Molecular Genetics, Department of Computational Molecular Biology, Berlin, Germany. [6]Biotech Research and Innovation Centre (BRIC), Department of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [7]Centre for Computational Biology, University of Birmingham, Birmingham, UK. [8]Biozentrum, University of Basel and Swiss Institute of Bioinformatics, Basel, Switzerland. [9]Max Planck Institute for Molecular Biomedicine, Muenster, Germany. [10]Translational and Functional Genomics Branch, National Human Genome Research Institute, Bethesda, MD, USA. [11]Southern Marine Science and Engineering Guangdong Laboratory, Guangzhou, China. [12]CAS Key Laboratory of Tropical Marine Bio-Resources and Ecology, South China Sea Institute of Oceanology, Chinese Academy of Sciences, Guangzhou, China. [13]Environmental Genomics & Systems Biology, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [14]Genomics and Epigenetics Division, Garvan Institute of Medical Research, Sydney, New South Wales, Australia. [15]Institute of Neuroscience, University of Oregon, Eugene, OR, USA. [16]Institute of Biological and Chemical Systems - Biological Information Processing (IBCS-BIP), Karlsruhe Institute of Technology, Karlsruhe, Germany. [17]Max-Delbrück-Center for Molecular Medicine in the Helmholtz Association (MDC), Berlin Institute for Medical Systems Biology (BIMSB), Berlin, Germany. [18]Sheffield Bioinformatics Core, Sheffield Institute of Translational Neuroscience, University of Sheffield, Sheffield, UK. [19]Singapore Institute for Clinical Sciences, Singapore, Singapore. [20]Bateson Centre/Biomedical Science, University of Sheffield, Sheffield, UK. [21]Centro Andaluz de Biología del Desarrollo (CABD), CSIC-Universidad Pablo de Olavide-Junta de Andalucía, Seville, Spain. [22]Epigenetics and Sex Development Group, Berlin Institute for Medical Systems Biology, Max-Delbrück Center for Molecular Medicine, Berlin, Germany. [23]International Institute of Molecular and Cell Biology in Warsaw, Warsaw, Poland. [24]MRC Weatherall Institute of Molecular Medicine, Radcliffe Department of Medicine, University of Oxford, Oxford, UK. [25]Vision Research Foundation, Sankara Nethralayas, Chennai, India. [26]Laboratory of Zebrafish Development and Disease Models (ZDDM), GIGA-R, SART TILMAN, University of Liège, Liège, Belgium. [27]Laboratory for Transcriptome Technology, RIKEN Center for Integrative Medical Sciences, Yokohama, Japan. [28]Institute of Biology, Humboldt University, Berlin, Germany. [29]Department of Developmental Biology, Signalling Research Centers BIOSS and CIBSS, University of Freiburg, Freiburg, Germany. [30]School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, New South Wales, Australia. [31]Randall Centre for Cell & Molecular Biophysics, Guy's Campus, King's College London, London, UK. [32]Science for Life Laboratory, Solna, Sweden. [33]Present address: Institute of Marine Sciences, Barcelona, Spain. [34]Present address: Fondazione Human Technopole, Milano, Italy. [35]These authors contributed equally: Damir Baranasic, Matthias Hörtenhuber, Piotr J. Balwierz, Tobias Zehnder, Abdul Kadir Mukarram. ✉e-mail: carsten.daub@ki.se; b.lenhard@imperial.ac.uk; f.mueller@bham.ac.uk

Methods

Resources and data availability. The resources produced by this publication, along with their location are as follows.

(1) Overview of the DANIO-CODE consortium and contributors (https://www.birmingham.ac.uk/generic/danio-code/index.aspx)
(2) DANIO-CODE Data Coordination Center (DCC) (danio-code.zfin.org)
(3) DANIO-CODE track hub for UCSC browser (danRer10: http://genome.ucsc.edu/cgi-bin/hgTracks?db=danRer10&hubUrl=https://danio-code.zfin.org/trackhub/DANIO-CODE.hub.txt danRer11: http://genome.ucsc.edu/cgi-bin/hgTracks?db=danRer11&hubUrl=https://danio-code.zfin.org/trackhub/DANIO-CODE.hub.txt)
(4) Session for WashU EpiGenome Browser (https://github.com/DANIO-CODE/DANIO-CODE_Data_analysis/tree/master/Figures/Figure1#figure-1c=)
(5) Motif Activity Response Analysis (MARA) (https://ismara.unibas.ch; DANIO-CODE results: https://ismara.unibas.ch/danio-code)
(6) Regulatory motifs and regulatory site annotations (https://swissregulon.unibas.ch/sr/downloads)
(7) Code repository for DANIO-CODE processing pipelines (gitlab.com/danio-code)
(8) Code repository for data analysis in this paper (https://github.com/DANIO-CODE/DANIO-CODE_Data_analysis)
(9) Videos with tutorials and example usages of the resource: https://youtube.com/playlist?list=PLiWQCe7dGqm6AtA0oP7qIaEQNa-7Z7fh5

Animal work. All animal work and associated methods are presented in the Supplementary Methods. Only early zebrafish embryos up to the free-feeding stage (5 days post fertilization) were used in this study. Zebrafish embryos/larvae up to the free-feeding stage are not considered as protected animals by law in the UK and are not subject to animal experimentation regulations. Breeding and maintenance of adult zebrafish strains was carried out in a designated facility under Home Office project licenses 40/3681 and P51AB7F76 assigned to the University of Birmingham, UK.

Data collection. We started the DANIO-CODE data collection aiming to capture a wide range of developing stages in zebrafish from a broad range of genomic, epigenomic and transcriptomic assays.

Members of the zebrafish community were invited to provide their published as well as unpublished data to the DANIO-CODE consortium. Benefiting from experiences of consolidating data in the decentralized data production of the modENCODE consortium[71], we developed the DANIO-CODE Data Coordination Center (DCC)[26] (https://danio-code.zfin.org). The DCC facilitated data collection as well as data annotation and subsequent data distribution.

Demultiplexed FASTQ files were provided by community members to the DCC file system. Using the DCC web frontend, the community members were guided through an annotation process to annotate the data they provided. The DCC data model is derived from Sequence Read Archive (SRA) data structures[72] and employs controlled vocabularies based on ZFIN nomenclature[73]. In addition to the community-provided data, DANIO-CODE annotators strategically selected additional published datasets to complement developmental stages or assays so far under-represented in the DCC. These datasets were annotated by the DANIO-CODE curators on the basis of the respective publications.

Consistent data and annotation formats allowed the consistent processing of all data in the DCC. For this, we developed computational workflows for all the data types and implemented these workflows for the DNAnexus system (https://www.dnanexus.com). Data and annotation quality control measures were established for all data in the DCC.

As a result, all datasets present in the DCC are described in terms of the overall study design, biosamples, library preparation methods, sequencing details as well as in data processing and quality control aspects. Snapshots of the DCC are kept as data freezes to facilitate the handling of newly added data. An interactive data and annotation view and export is provided at https://danio-code.zfin.org/dataExport.

Transcripts identification. Wild-type embryonic paired-end and stranded RNA-seq samples (DCD000141SR, DCD000225SR, DCD000247SR, DCD000433SR, DCD000426SR, DCD000324SR)[31,32,74–77] were selected from a total of 528 DANIO-CODE RNA-seq and aligned to GRCz10 using STAR aligner v.2.5.1b (ref. [78]). StringTie v.1.33b (ref. [79]) was used to call transcripts, which were then all assembled using TACO[80], generating a total of 194,508 transcripts in canonical chromosomes. Transcript quantification was done using Salmon v.0.11.2 (ref. [81]). We removed read through, mono-exonic and transcripts overlapping three or more Ensembl genes. All protein-coding transcripts above 200 kb and long noncoding RNAs above 100 kb were excluded (permissive set). We defined transcripts that are expressed in a minimum of two closest stages as the robust set. To get high confidence transcripts, we added those that have consensus CAGE transcription start site clusters (TCs) in the proximity (±500 bp), yielding 70,354 transcripts (permissive set) and 55,596 transcripts (robust set). Out of 35,117 Ensembl genes (v.91), 22,065 and 23,568 genes were covered in our robust and permissive sets, respectively.

Promoterome construction. First, all reads mapping to poorly assembled genomic regions or otherwise blacklisted[82] regions were excluded from the set of CAGE-supported TSSs (cTSSs). After an initial application of CAGEr[83] we discovered systematic differences between nAnTi and tagging CAGE samples both at the number of transcription start site clusters number and summed promoter/gene expression across samples produced with the two CAGE protocols. In particular, the fraction of CAGE signal coming from annotated exons was elevated in nAnTi samples, skewing the statistics. To a varying extent this phenomenon (known as exon painting/carpeting) has been previously observed and attributed to recapping of degraded mRNAs. Since the majority of true TSSs are initiated at either YC or YR dinucleotides[9], we analyzed dinucleotide frequencies at initiation sites and confirmed an increased proportion of other (non-YC and non-YR) dinucleotides in nAnTi compared to tagging samples. We therefore decided to remove all CAGE tags not initiated at YC or YR dinucleotides.

The remaining set of TSSs was power-law normalized[84] to a common exponent alpha = 1.1 and 5 to 1,000 tags fit range, and the TCs were produced using the following parameters of CAGEr: threshold = 0.7, thresholdIsTpm = TRUE, nrPassThreshold = 1, method = 'distclu', maxDist = 20, removeSingletons = TRUE, keepSingletonsAbove = 5. This yielded a comparable number of TCs across samples without an obvious bias towards high numbers in nAnTi samples. The number of TCs moderately increased in post-ZGA samples.

To compare expression levels across samples we called consensus clusters (genomic regions not assigned to any particular sample, unlike TCs) with settings tpmThreshold = 1.0, qLow = NULL, qUp = NULL, maxDist = 20. To further filter weak or spurious tag clusters we kept consensus clusters that were expressed in at least two consecutive developmental stages. Specifically, we required that there exist a TC within a consensus cluster in both consecutive stages with at least 1.0 tags per million (tpm) expression. This yielded 27,781 consensus clusters. We calculated expression of each consensus cluster by summing all YC- and YR-initiated CAGE tags from within the cluster across stages. This approach differs from CAGEr implementation, which includes expression only from TCs within a consensus cluster and is subject to generating noise at lowly expressed regions due to the tpmThreshold parameter.

To visualize obtained expression levels we made a two-dimensional principal components analysis plot, which correctly grouped nAnTi CAGE and tagging CAGE samples from the same stage.

Annotation of alternative transcript and alternative promoter. A gene can have multiple transcripts/isoforms that differ in their TSSs by few nucleotides, to tens of kilobases. When a gene has multiple transcripts, Ensembl assigns the longest transcript as a reference transcript and its promoter as a reference promoter. To comprehensively assign CAGE peaks to transcripts, we analyzed transcript models from Ensembl, RefSeq and novel RNA-seq transcripts from DANIO-CODE. Thus, we focused only on the transcripts that are supported by CAGE peaks. Similar to the Ensembl method, we annotated the longest transcript as the reference transcript and its promoter as the reference promoter. Remaining transcripts whose TSSs were proximal (<300 nucleotides) to the assigned reference transcripts were excluded. On the remaining distal transcripts, the longest transcript was assigned as an alternative transcript and excluded other transcripts with proximal (<300 nucleotides) transcription start sites. Some genes have more than one alternative transcript, thus we iterated this process to annotate additional alternative promoters that are distal from assigned reference or alternative transcripts.

To annotate alternative promoters utilized during mouse embryonic developmental stages, we analyzed FANTOM5 CAGE-seq data[28] from four embryonic stages (E11 days, E12 days, E13 days and E14 days), which are similar to the zebrafish stages analyzed. We analyzed Ensembl, RefSeq and RNA-seq transcripts models downloaded from UCSC table browser. We used a similar method to that described above to annotate alternative promoters in mouse. To identify orthologs of alternative transcripts/promoters, we downloaded mouse/zebrafish ortholog tables from Ensembl[85].

Motif activity analysis. To curate a set of regulatory motifs for zebrafish we first collected all Pfam models that corresponded to DNA-binding domains (DBDs). To define a set of zebrafish TFs we extracted a representative protein sequence for each zebrafish gene, ran HMMER with these Pfam models and extracted the DBD sequences for each protein with substantial hits. Starting from a previously curated collection of regulatory motifs for human and mouse[36], we extracted the DBD sequences of the human and mouse TFs associated with these regulatory motifs. Using BLAT[86], we then aligned all zebrafish DBD sequences to the human/mouse DBD sequences and associated zebrafish TFs with the human or mouse TF (and regulatory motif) that best matched their DBD sequences. Note that multiple zebrafish TFs can thus end up being associated with the same regulatory motif. These procedures led to 814 zebrafish TF genes being assigned to 581 unique regulatory motifs.

For each promoter, we defined the proximal promoter region as the region from 500 base pairs (bp) upstream to 500 base pairs downstream of the CAGE transcription start region (TSR). For each proximal promoter region, we obtained the orthologous regions from the goldfish, common carp and grass carp genomes using LAST[87] and multiply aligned the orthologous regions using T-coffee[88]. We

also obtained a phylogenetic tree of the four species from the observed fractions of conserved nucleotides in the promoter alignments of each pair of species. For each regulatory motif, we then ran MotEvo[89] on these multiple alignments to obtain TFBS predictions genome wide. Using these TFBS predictions, we constructed a site count matrix N for the MARA analysis, where each component N_{pm} corresponds to the sum of the posterior probabilities of all binding sites for motif m in promoter p. Motifs whose site counts across promoters genome wide had a higher Pearson correlation than $r=0.6$ were grouped into motif groups, leading to 489 motif groups. ISMARA analysis was then performed on the CAGE expression data across the developmental time course[89].

Functional segmentation of the genome. We identified *cis*-regulatory elements genome wide using their characteristic ChIP–seq signal. For example, acetylation of lysine residue 27 and monomethylation of lysine residue 4 on the histone H3 (H3K27ac and H3K4me1) are features of active chromatin. The modification H3K4me3 is characteristically found on promoters, while H3K27me3 represents Polycomb-repressed chromatin. Those modifications were localized ChIP–seq. We used ChromHMM[43,44] to segment the genome into regions containing specific chromatin marks. We captured the epigenetic state of the genome in five different development stages. We used published data for the Dome, 75% epiboly, 5–9 somites, Prim-5 and Long-pec stages[4,31,90–93], as well as newly produced data for the 5–9 somites and Long-pec stages. After optimization, we found ten optimal latent states on basis of the emission parameters of chromatin marks. The states were matched between stages and manually assigned a function using The Roadmap Epigenome Project[21] annotation as a reference. The identified functional elements were annotated as follows:

1. Active TSS 1 (1_TssA1), 2. Active TSS 2 (2_TssA2), 3. TSS Flanking region 1 (3_TssFlank1), 4. TSS Flanking region 2 (4_TssFlank2), 5. Active enhancer 1 (5_EnhA1), 6. Enhancers flanking region (6_EnhFlank), 7. Primed enhancer (7_EnhWk1), 8. Poised elements (8_Poised), 9. Polycomb-repressed regions (9_PcRep), 10. Quiescent state (10_Quies).

The active promoters and promoter flanking regions, in addition to active chromatin marks, show strong emission of H3K4me3. Moreover, the promoter-associated states are mostly found on and around the annotated TSS. The states missing H3K4me3 and not found around TSSs were annotated as enhancer related. Depending on whether both H3K27ac and H3K4me1 are present, as well as the strength of the emission, the enhancer states were divided into active enhancers (strong H3K27ac and H3K4me1 emission, but no H3K4me3), enhancers flanking (weak H3K27ac emission, mostly found around active enhancers) and primed enhancers (H3K4me1 emission only). States emitting H3K27me3 were annotated as Polycomb related. In addition to H3K27me3, when active marks were present, the state was assigned as poised; otherwise, it was assigned as Polycomb repressed. When no marks were present, the region was assigned as quiescent. Most of the genome shows no marks at all.

PADREs. We constrained our subsequent analyses to the regions in the genomes that are open, that is, depleted in nucleosomes as identified by ATAC-seq. We identified stage-specific open chromatin regions consistent between replicates, with the irreproducibility discovery rate[94] less than 0.1, in seven developmental stages (four pre-ZGA stages, newly produced datasets and seven post-ZGA stages, of which 30% epiboly is newly produced and the other samples were published previously[77]). We termed those regions as predicted ATAC-supported developmental regulatory elements. The reason for naming them in this way was to distinguish them from ENCODE cCREs in two segments as: (1) they contain open regions even without the support of functional marks, and (2) we wanted to emphasize the developmental aspect of the defined set of elements. All stage-specific PADREs were merged to form a set of regions called consensus PADREs (cPADREs). Two different sets of cPADREs were defined. The permissive cPADREs consist of all PADREs merged and number around ~240, 000 elements. The strict set considers regions that are open in at least two neighboring stages. This set counts ~140,000 elements. All cPADRE analyses in this paper were done on strict cPADREs. We assigned each ATAC-seq peak a functional annotation on the basis of overlaps with the ChromHMM state in available stages.

UMAP visualization. We developed a method that considers various signals around the open chromatin summit comprehensively. In brief, we constructed a feature matrix using ATAC-seq, H3K4me3, H3K27ac, H3K4me1 ChIP–seq tags, as well as nucleosome position calculated by NucleoATAC[95] (Extended Data Fig. 6a–c). Nucleosome signal was included because some factors have well-positioned nucleosomes around their binding sites and could separate those factors from others. In brief, the peak summit is extended for 750 bp in each direction and split into 13 bins (R1–R13). For each bin, the number of tags for the aforementioned assay types is counted, and the mean nucleosome signal in each bin was calculated using the genomation package. This resulted in five score matrices, each having the number of rows the same as the number of open chromatin regions and 13 columns (one for each bin around the peak summit). Those matrices were standardized by scaling the values and centering the mean to 0. The standardized matrices were concatenated column-wise, giving a total of 65 columns. Using the UMAP algorithm[96], the number of features was reduced from 65 to 2, making it possible to plot each open chromatin region in a two-dimensional plot.

For the conservation analyses, the cyprinid (grass carp, common carp, goldfish and zebrafish) phastCons scores from Chen et al.[97] were used.

COPEs and DOPEs. Constitutive elements were defined as the intersection of distal PADREs at every developmental stage. COPEs were defined as constitutive, annotated as quiescent at every developmental stage. DOPEs were defined as cPADREs, annotated as quiescent at every developmental stage. DOPEs were further classified as adults-marked DOPEs if they overlapped H3K27ac marked regions in any of the adult tissues[24].

Promoter classification by open chromatin. For each TSR we defined a reference point as the TSS with the highest mean post-MBT expression as 'dominant TSS' (tpm values of samples ranging from the Shield to Long-pec stages) and required that it amounts to at least 0.2 tpm. This further reduced the set of consensus clusters to 21,914 elements. We then merged ATAC samples from the Prim-5 stage and extracted Tn5 cut sites from both ends of ATAC reads while correcting for Tn5 overhang, smoothed with a Gaussian kernel with standard deviation 3 bp and log-transformed. These ATAC cut-site profiles served as input to k-means clustering ($k=8$, range ±800 bp from the dominant TSS).

H3K27ac ensemble identification. Enhancer ensembles were detected using H3K27ac peaks and mapped reads from the Dome stage (DCC data identities: DCD006167DT and DCD008973DT) as input for the ROSE algorithm[57] with the distance from TSS to exclude adjusted to 500 bp.

Genomic coordinate projection. Genomic coordinates of GRB loci were projected between zebrafish and mouse using multiple pairwise sequence alignments between a set of 15 species. The basic concept of our approach is that, under the assumption of conserved synteny, a nonalignable genomic region can be projected from one species to another by interpolating its relative position between two alignable anchor points. The accuracy of such interpolations correlates with the distance to an anchor point. Therefore, projections between species with large evolutionary distances, such as zebrafish and mouse, tend to be inaccurate due to a low anchor point density. Including so-called bridging species may increase the anchor point density and thus improve projection accuracy. Extended Data Figure 10b illustrates the potential benefit of using a bridging species, with a schematic example projection between zebrafish and mouse. The optimal choice of bridging species may vary between different genomic locations and there may be genomic locations for which a combination of bridging species with intermediate projections produces optimal results. Extended Data Figure 10c presents the bridging species optimization problem as the shortest path problem in a graph where every node is a species and the weighted edges between them represent the distance of a genomic location to its anchor point. For that, we established a scoring function that reflects those distances and returns values between 0 and 1, where a score of 1 means that a genomic location x overlaps an anchor point a. The score decreases exponentially as the distance $|x-a|$ increases. For a single species comparison, the function is defined as follows:

$$f(x_i) = \exp\left(-\frac{\min\left(\left|x_i - a_i^{(1)}\right|, \left|x_i - a_i^{(2)}\right|\right)}{g_i s}\right),$$

with g denoting the genome size of the respective species and s a scaling factor that can be determined by defining a distance half-life d_h, as the distance $|x-a|$ at which the scoring function returns a value of 0.5:

$$s = -\frac{d_h}{g\log(0.5)}.$$

The length of a path through the graph is then given by subtracting the product of the distance scoring function for every node in the path from 1:

$$l_p = 1 - \prod_{i\in p} f(x_i).$$

The shortest path $\dot{p}$ through the graph is then found by minimizing l_p:

$$\dot{p} = \arg\min_{p\in P} l_p,$$

with P denoting the set of all paths through the graph. The optimization problem presents a classic shortest path problem and is solved using Dijkstra's shortest path aAlgorithm[98].

Epigenomic profile comparison. We compared H3K27me3 ChIP–seq data from phylotypic stages in zebrafish (Prim-5 stage) and mouse (E10.5 stage)[99], when their transcriptomes are most similar[100]. To match the whole-embryo zebrafish data, we created a virtual embryo dataset for mouse by merging data for six different tissues

(fore-, mid-, hindbrain, facial prominence, heart, limb). The mouse H3K27me3 profile was projected onto zebrafish genomic coordinates using the multispecies approach by splitting the zebrafish GRB into 1-kb windows, projecting their center coordinates onto mouse and retrieving the signal from the respective 1-kb bin in mouse. 'Signal' stands for H3K27me3 coverage represented as quantiles after quantile normalization of the two distributions in zebrafish and mouse. Signal overlap is represented by the log signal ratio and capped to values in [−1,1]. Signal amplitude represents the maximum signal of zebrafish and mouse to the power of 10 to increase the variance of signal amplitude. For mouse and zebrafish comparison of H3K27ac ensembles, previously published data were used[101].

Classification of conservation. Zebrafish ATAC-seq peaks were classified into three conservation classes on the basis of the projection using the multispecies approach. Directly conserved ATAC-seq peaks overlap a direct alignment between zebrafish and mouse. Indirectly conserved ATAC-seq peaks do not overlap a direct alignment, but are projected with a score >0.99, that is, either overlapping or very close to a multispecies anchor. The remaining peaks are classified as nonconserved. A score of 0.99 means that the sum of the distances from peak to anchor points is <150 bp considering all intermediate species in the optimal species path.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Raw and aligned sequencing data are available at https://danio-code.zfin.org/dataExport/. The raw sequencing data produced for this study are available on the European Bioinformatics Institute (EBI) European Nucleotide Archive (ENA) under study numbers PRJNA824720, PRJNA821001, PRJNA821088, PRJNA821148 and PRJNA821034. Annotation tracks are available at http://genome.ucsc.edu/cgi-bin/hgTracks?db=danRer10&hubUrl=https://danio-code.zfin.org/trackhub/DANIO-CODE.hub.txt (danRer10) and http://genome.ucsc.edu/cgi-bin/hgTracks?db=danRer11&hubUrl=https://danio-code.zfin.org/trackhub/DANIO-CODE.hub.txt (danRer11).

Code availability

The processing pipelines for the individual assays are available at https://gitlab.com/danio-code. The code used for the analysis is available at https://github.com/DANIO-CODE/DANIO-CODE_Data_analysis (https://doi.org/10.5281/zenodo.6424702). The script to generate the TrackHub is available at https://gitlab.com/danio-code/TrackHub.

References

71. Celniker, S. E. et al. Unlocking the secrets of the genome. *Nature* **459**, 927–930 (2009).
72. Kodama, Y., Shumway, M. & Leinonen, R. International Nucleotide Sequence Database Collaboration The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.* **40**, D54–D56 (2012).
73. Ruzicka, L. et al. The Zebrafish Information Network: new support for non-coding genes, richer Gene Ontology annotations and the Alliance of Genome Resources. *Nucleic Acids Res.* **47**, D867–D873 (2019).
74. Lee, M. T. et al. Nanog, Pou5f1 and SoxB1 activate zygotic gene expression during the maternal-to-zygotic transition. *Nature* **503**, 360–364 (2013).
75. Etard, C. et al. Loss of function of myosin chaperones triggers Hsf1-mediated transcriptional response in skeletal muscle cells. *Genome Biol.* **16**, 267 (2015).
76. Meier, M. et al. Cohesin facilitates zygotic genome activation in zebrafish. *Development* **145**, dev156521 (2017).
77. Marlétaz, F. et al. Amphioxus functional genomics and the origins of vertebrate gene regulation. *Nature* **564**, 64–70 (2018).
78. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
79. Pertea, M., Kim, D., Pertea, G. M., Leek, J. T. & Salzberg, S. L. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nat. Protoc.* **11**, 1650–1667 (2016).
80. Niknafs, Y. S., Pandian, B., Iyer, H. K., Chinnaiyan, A. M. & Iyer, M. K. TACO produces robust multisample transcriptome assemblies from RNA-seq. *Nat. Methods* **14**, 68–70 (2016).
81. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).
82. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE blacklist: identification of problematic regions of the genome. *Sci. Rep.* **9**, 9354 (2019).
83. Haberle, V., Forrest, A. R. R., Hayashizaki, Y., Carninci, P. & Lenhard, B. CAGEr: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. *Nucleic Acids Res.* **43**, e51 (2015).
84. Balwierz, P. J. et al. Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. *Genome Biol.* **10**, R79 (2009).
85. Zerbino, D. R. et al. Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).
86. Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
87. Frith, M. C. & Kawaguchi, R. Split-alignment of genomes finds orthologies more accurately. *Genome Biol.* **16**, 106 (2015).
88. Notredame, C., Higgins, D. G. & Heringa, J. T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205–217 (2000).
89. Arnold, P., Erb, I., Pachkov, M., Molina, N. & van Nimwegen, E. MotEvo: integrated Bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences. *Bioinformatics* **28**, 487–494 (2012).
90. Irimia, M. et al. Extensive conservation of ancient microsynteny across metazoans due to *cis*-regulatory constraints. *Genome Res.* **22**, 2356–2367 (2012).
91. de la Calle Mustienes, E., Gómez-Skarmeta, J. L. & Bogdanović, O. Genome-wide epigenetic cross-talk between DNA methylation and H3K27me3 in zebrafish embryos. *Genomics Data* **6**, 7–9 (2015).
92. Nepal, C. et al. Dynamic regulation of the transcription initiation landscape at single nucleotide resolution during vertebrate embryogenesis. *Genome Res.* **23**, 1938–1950 (2013).
93. Ulitsky, I., Shkumatava, A., Jan, C. H., Sive, H. & Bartel, D. P. Conserved function of lincRNAs in vertebrate embryonic ddespite rapid sequence evolution. *Cell* **147**, 1537–1550 (2011).
94. Li, Q., Brown, J. B., Huang, H. & Bickel, P. J. Measuring reproducibility of high-throughput experiments. *Ann. Appl. Stat.* **5**, 1752–1779 (2011).
95. Schep, A. N. et al. Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. *Genome Res.* **25**, 1757–1770 (2015).
96. McInnes, L., Healy, J., Saul, N. & Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *J. Open Source Softw.* **3**, 861 (2018).
97. Chen, Z. et al. De novo assembly of the goldfish (*Carassius auratus*) genome and the evolution of genes after whole-genome duplication. *Sci. Adv.* **5**, eaav0547 (2019).
98. Dijkstra, E. W. A note on two problems in connexion with graphs. *Numer. Math.* **1**, 269–271 (1959).
99. Gorkin, D. U. et al. An atlas of dynamic chromatin landscapes in mouse fetal development. *Nature* **583**, 744–751 (2020).
100. Irie, N. & Kuratani, S. Comparative transcriptome analysis reveals vertebrate phylotypic period during organogenesis. *Nat. Commun.* **2**, 248 (2011).
101. Zhang, T., Zhang, Z., Dong, Q., Xiong, J. & Zhu, B. Histone H3K27 acetylation is dispensable for enhancer activity in mouse embryonic stem cells. *Genome Biol.* **21**, 45 (2020).

Acknowledgements

We are indebted to the late Jose Luis Gomez Skarmeta for his devoted support of the DANIO-CODE programme. We thank M. Haussler at UCSC and D. Zerbino at EBI for facilitating access of DANIO-CODE track hubs in the UCSC and Ensembl genome browsers, respectively. We thank ZFIN for hosting the DANIO-CODE DCC and raw data. We thank J. Horsefield for creating the DANIO-CODE logo. We thank data producers (for the list of laboratories visit the DANIO-CODE DCC) who directly uploaded data and provided metadata directly. We thank DNANexus for providing computer time for the reprocessing of public datasets. We thank our main funders, the Horizon 2020 MSCA-ITN project ZENCODE-ITN by the European Commission to F.M., B.L., C.O.D., J.M.V. and B.P. (GA no: 643062), BBSRC support (DanioPeaks, P61715) to B.L., F.M. and F.C.W., Wellcome Trust (Joint-Investigator award 106955/Z/15/Z) to F.M. and B.L. and AQUA-FAANG (Horizon 2020, GA 817923) to B.L., D.B. and F.M. and BBSRC (BB/R015457/1) to F.vE and Key Special Project for Introduced Talents Team to Z.C. (GML2019ZD0401) and PrecisionTox project by the European Commission (GA no: 965406). We thank SNP&SEQ Technology Platform in Uppsala, Sweden (CAGE sequencing), MRC LMS Genomics Facility and Genomics Birmingham facilities UK. D.B. was awarded the Rutherford Fund Fellowship.

Author contributions

The DANIO-CODE DCC implementation was led by M.H. together with A.K.M. and M.S. Data curation was done by M.H., M.D. and I.S. Data processing was done by D.B., M.H., A.K.M., P.J.B., B.H.-R., M.D., I.S., S.E.R. and R.D.A. Data analysis was done by D.B., M.H., P.J.B., T.Z., A.K.M., C.N., C.V. and S.E.R. Data visualization was done by M.H. New samples were prepared by Y.H., A.J.-G., N.L., J.W., F.M.D., N.D., R.D.A. and S.N. Track hubs were generated by D.B., M.H. and Z.C. ZFIN support was provided by A.E. and R.M. Motif activity analysis was implemented by D.R. and M.P., and analyzed by P.J.B., D.R., M.P. and E.v.N. Transgenically validated enhancer elements list was generated and curated by O.O., Y.H. and S.R.. Cell-type specificity from single-cell ATAC-seq data were provided by A.C.M. and W.K. dCAS9 experiments were performed by E.C., D.W. and H.R.K. Unpublished data were provided by M.L., V.C., S.M., B.P., T.S.-S., C.W.,

J.M.V., D.O., O.B., F.C.W. and F.M. M.V., P.C., U.O., S.A.L., S.M.B., C.W., F.v.E., J.M.V., D.O., B.J.B., O.B., M.W., F.C.W., C.O.D., B.L. and F.M. supervised the analyses and provided feedback. The manuscript was written by D.B., M.H., P.J.B., T.Z., A.K.M., C.N., S.R., C.W., D.O., E.v.N., F.C.W., C.O.D., B.L. and F.M., with other authors contributing useful comments and feedback. Data analysis planning and coordination was done by D.B., C.O.D., B.L. and F.M. The DANIO-CODE initiative was conceptualized by F.C.W., B.L. and F.M. and coordinated by F.M.

Competing interests

The authors declare no competing interests.

Additional information

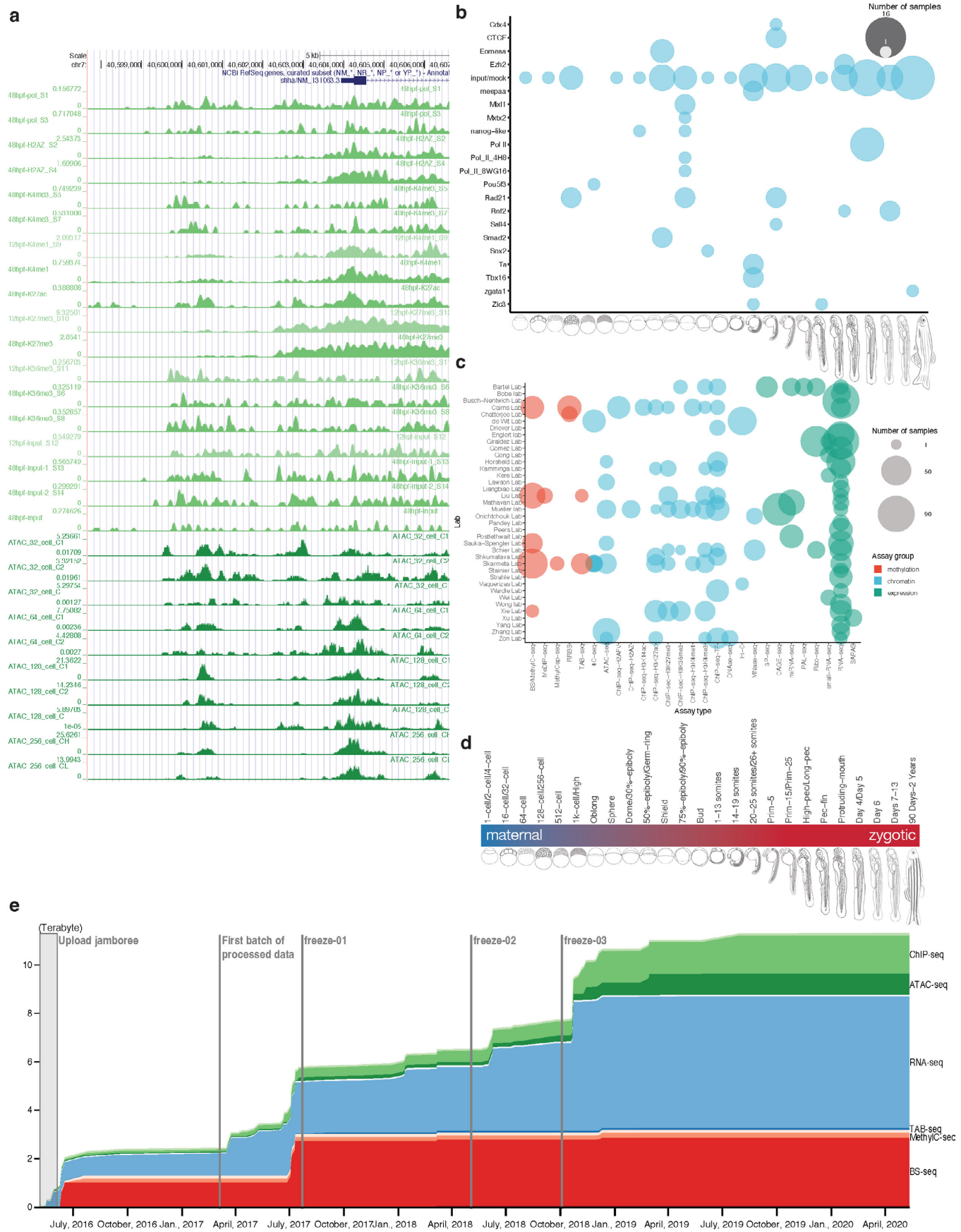
Extended data is available for this paper at https://doi.org/10.1038/s41588-022-01089-w.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-022-01089-w.

Correspondence and requests for materials should be addressed to Carsten O. Daub, Boris Lenhard or Ferenc Müller.

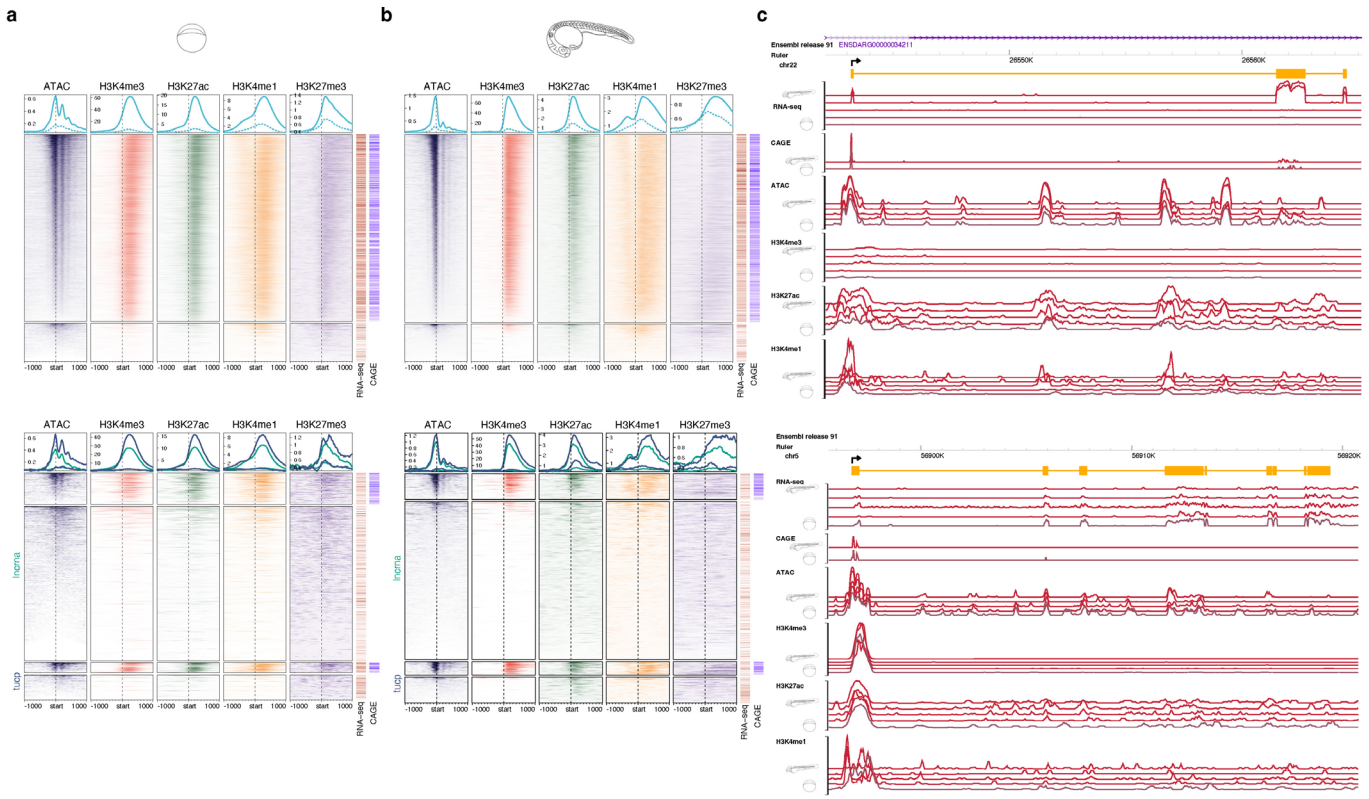
Peer review information *Nature Genetics* thanks Roderic Guigó and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

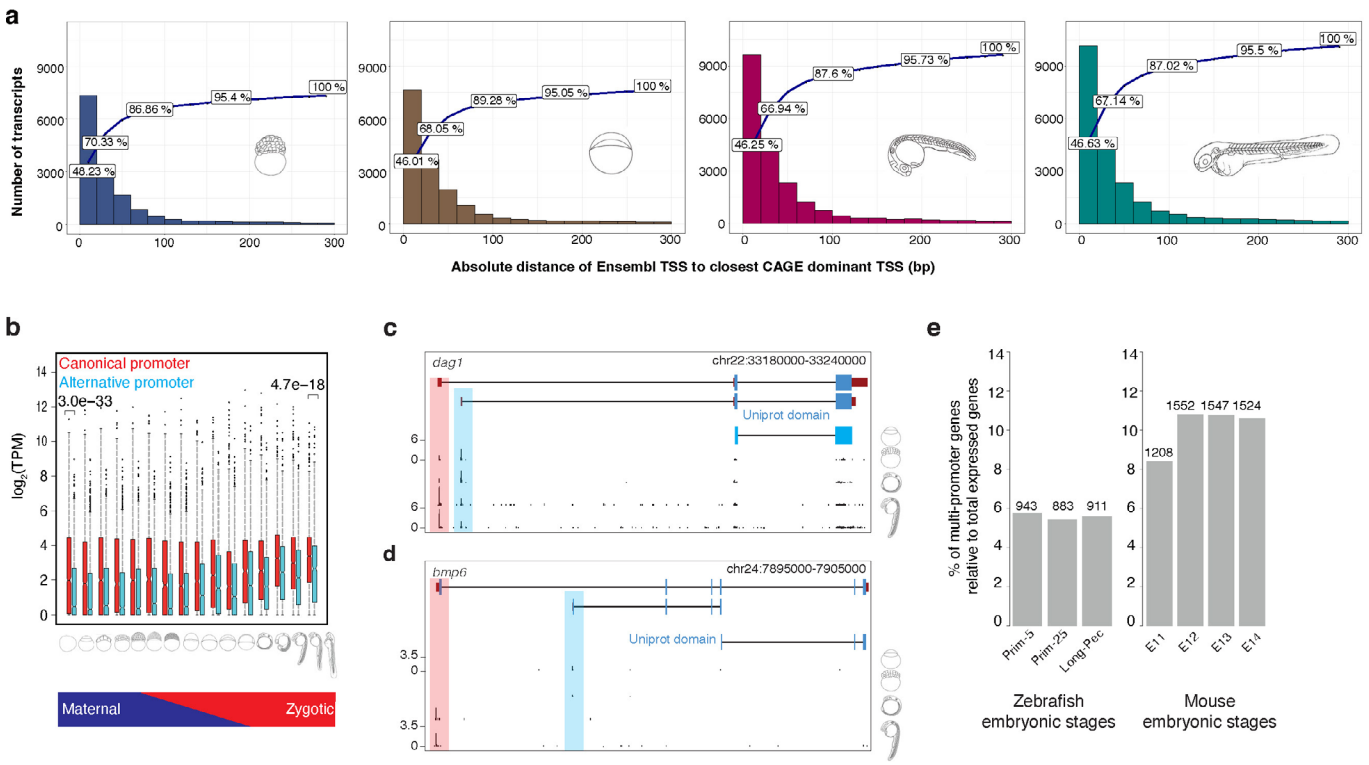


Extended Data Fig. 1 | See next page for caption.

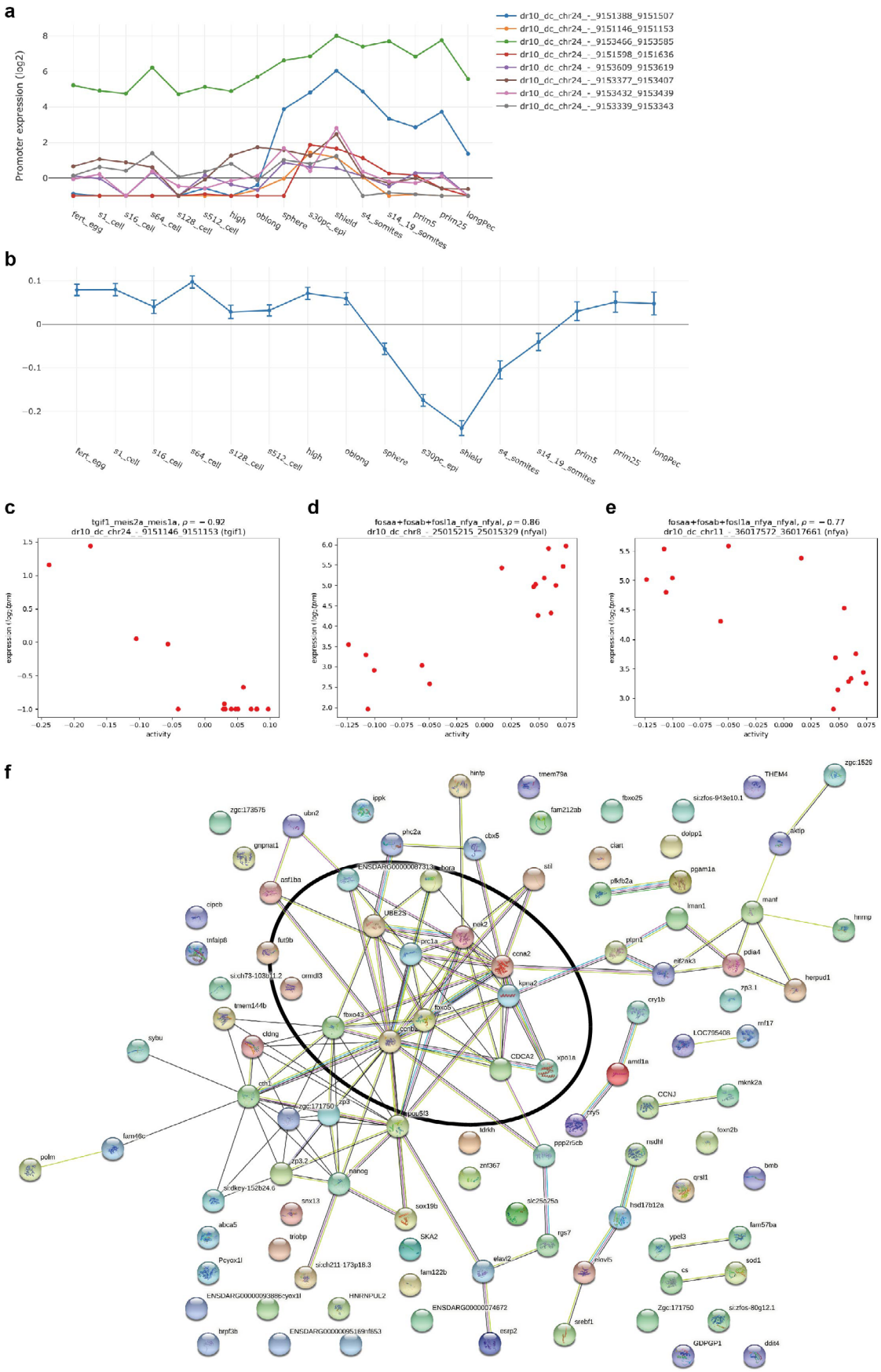
Extended Data Fig. 1 | Data increase in the DANIO-CODE Data Coordination Center. a, Tracks of representative examples of unpublished datasets in a UCSC Genome Browser session including CAGE, ATAC, and ChIP datasets generated by DANIO-CODE laboratories. Promoter region of developmental regulator *shha* gene is shown. **b**, DCC Data availability summary for ChIP with antibodies against Pol II, CTCF and transcription factors as indicated. Stages and stage ranges are indicated on the X axis, the transcription factor occupancy detected is listed on the Y axis. **c**, Data producers and data types matrix indicating the data producer lab (Y axis) and the type of data (X axis). **d**, Data acquisition evolution in the DCC.



Extended Data Fig. 2 | Validation of annotated transcripts. a, b, Aggregation plots and heatmaps of open chromatin and epigenetic features of annotated transcripts and CAGE-seq validation of TSSs (bars on the right of each panel) are shown for the Dome (**a**) and Prim-5 (**b**) stages. Top panels show protein coding genes (n=14,471, of which 12,031 are supported by CAGE for the Dome stage; n=16,478, of which 13,769 are supported by CAGE for the Prim-5 stage) and bottom panels show lncRNA (n=1,780, of which 302 are supported by CAGE for the Dome stage; n=1,551, of which 220 are supported by CAGE for prim-5) and TUCP (n=336, of which 97 are supported by CAGE for the Dome stage; n=329, of which 112 are supported by CAGE for the Prim-5 stage) genes **c**, Example screenshot of novel lncRNA(top), and novel TUCP (bottom) transcripts and associated epigenomic features.

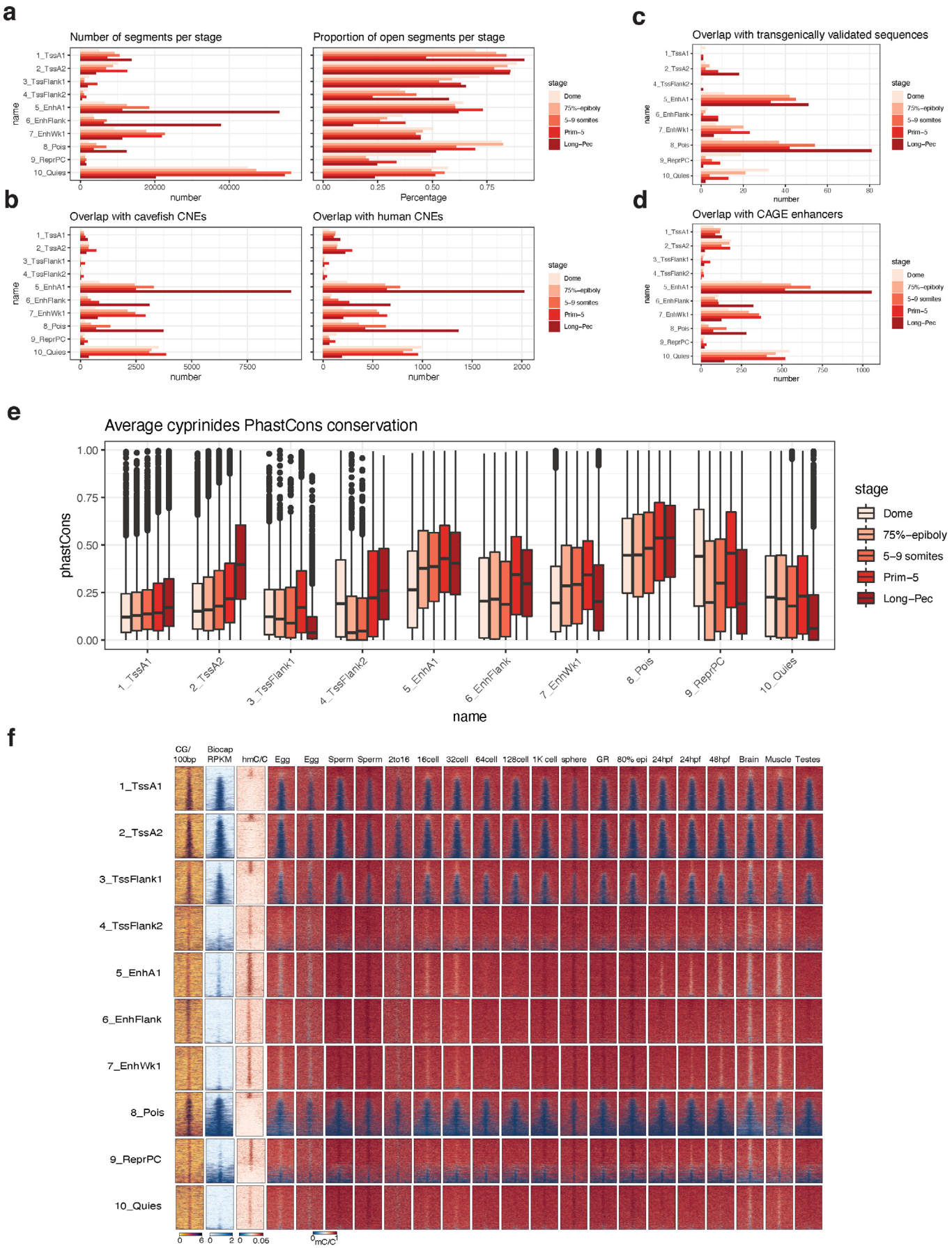


Extended Data Fig. 3 | Characterisation of promoter-calling precision and alternative promoter usage in annotated transcripts. a, Frequency distribution of Ensemble transcripts 5′ ends binned according to distance (bp) from CAGE dominant peak as indicated on X axis. Cumulative frequency depicted by line. Developmental stages are indicated by embryo symbols. **b**, Box plot shows the expression levels of canonical and alternative promoters across 16 developmental stages. P-values denote the significant difference in expression levels between canonical and alternative promoters during two stages at fertilized-egg (P=3.0E-33; t-test two-sided) and long pec (P=4.7E-18; t-test two-sided). **c**, A UCSC browser screenshot of the gene dag1 shows the alternative promoter (highlighted in cyan) is upstream of the start codon (pointed by arrow), thus altering only 5′UTR but not protein. The numbers on the y-axis represent the normalized tags per million (TPM) of CAGE tags. The Uniprot domain track denotes the annotated protein domain in the Uniprot database. **d**, A UCSC browser screenshot of the gene bmp6 shows the alternative promoter (highlighted in cyan) is downstream of the start codon (pointed by arrow) and alters the N terminal of the protein. **e**, Bar plots show the fraction of multi-promoter genes relative to the total expressed genes in zebrafish and mouse embryonic stages. The numbers on top of bar plot represent the actual number of multi-promoter genes. E11 represents embryonic day 11 and so on for E12, E13 and E14.



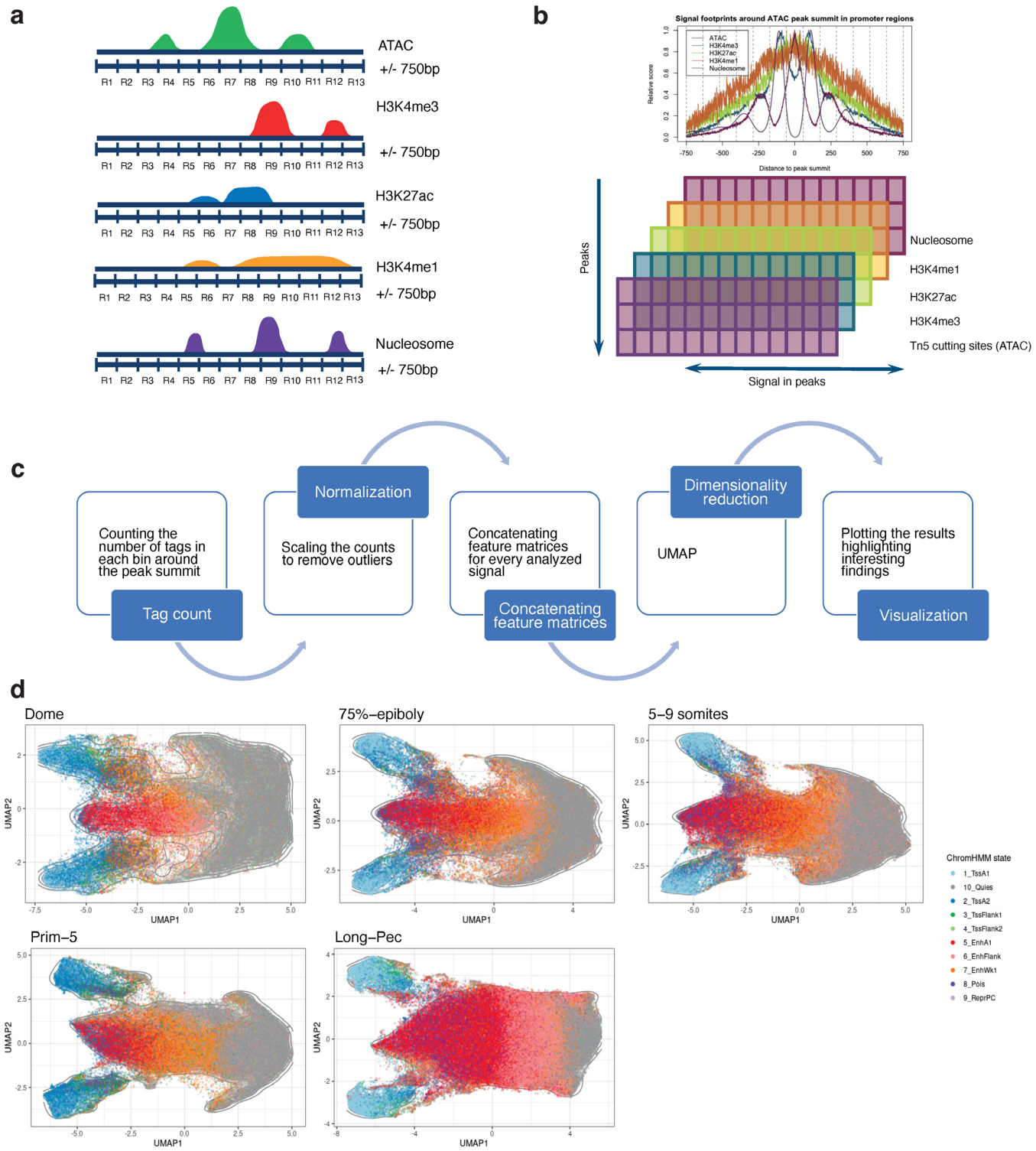
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Motif activity analysis. MARA predicts up-regulation of Tead3's activity from gastrulation onwards (Fig. 2). For each potential target promoter with Tead3 binding site, MARA quantifies the extent the Tead3 motif activity explains the target's expression dynamics (log-likelihood score). For each GO category the sum of log-likelihoods for all genes in the category was calculated. Supplementary Table 13 shows the GO biological process categories with the highest total log-likelihoods. Top categories correspond to processes in which Hippo signalling during early development in zebrafish has been implicated. **a,b**, The *tgif1* promoters are transiently upregulated during gastrulation **(a)** while the targets of Tgif1 are transiently down-regulated (**b**), supporting Tgif1 as a repressor. Posterior means and standard deviations (depicted as error bars) are based on analysis of the expression levels of all n= 27781 promoters for each sample. **c**, Scatterplot of TGIF1_MEIS1a_MEIS2a motif activity (horizontal axis) against total *tgif1* mRNA expression (vertical axis) shows motif activity and TF expression are highly anti-correlated (Pearson correlation -0.92). **d**, Scatterplot of the FOS/NF-Y motif activity (horizontal axis) against expression of the *nfyal* gene shows positive correlation (Pearson correlation coefficient 0.86). **e**, Scatterplot of the FOS/NF-Y motif activity (horizontal axis) against expression of the *nfya* gene shows negative correlation (Pearson correlation -0.77). As shown in Fig. 3d, MARA predicts that targets of NF-Y are down-regulated from the sphere stage onwards, thus as the NF-Y motif activity decreases during development. The expression of *nfya* is up whereas *nfyal* is down-regulated, suggesting that Nfya may replace Nfyal in the NF-Y complex. **f**, STRING database network picture of the predicted target genes of the NF-Y motif. The black oval indicates a set of target genes involved in mitosis and G2/M transition, consistent with the documented role of NF-Y.

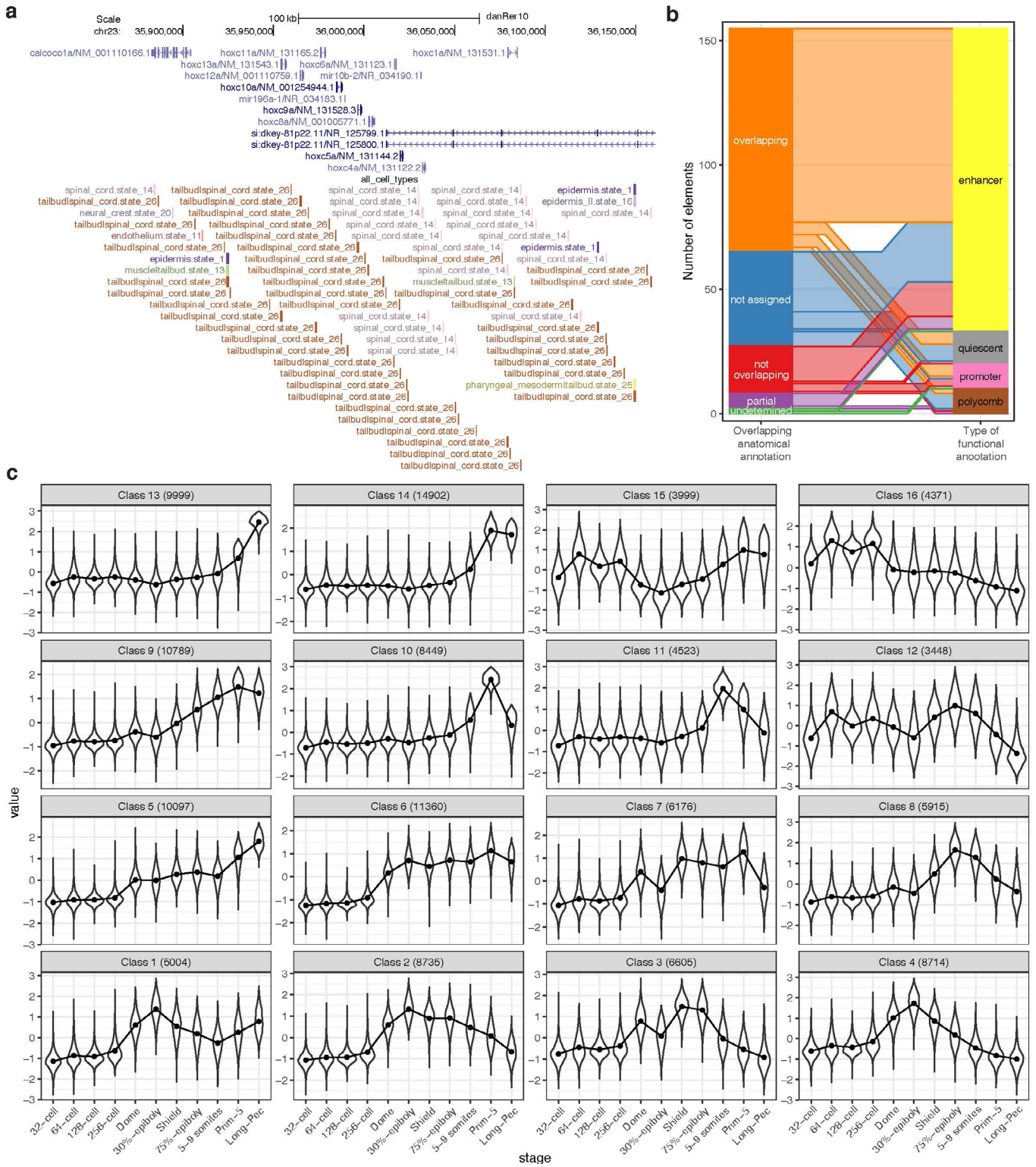


Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | PADREs validation. a, Left: Number of PADREs assigned to each chromatin state for every developmental stage. Right: Proportion of ChromHMM states present in PADREs for each stage. **b**, Number of annotated PADREs overlapping Mexican cavefish (left) and human (right) CNEs for each stage. **c**, Number of annotated PADREs overlapping transgenically validated enhancers for each stage. **d**, Number of annotated PADREs overlapping CAGE-defined eRNAs for each stage. **e**, phastCons scores distribution of annotated PADREs for each stage. **f**, Methylation profile throughout the development of annotated PADREs at the Prim-5 stage.

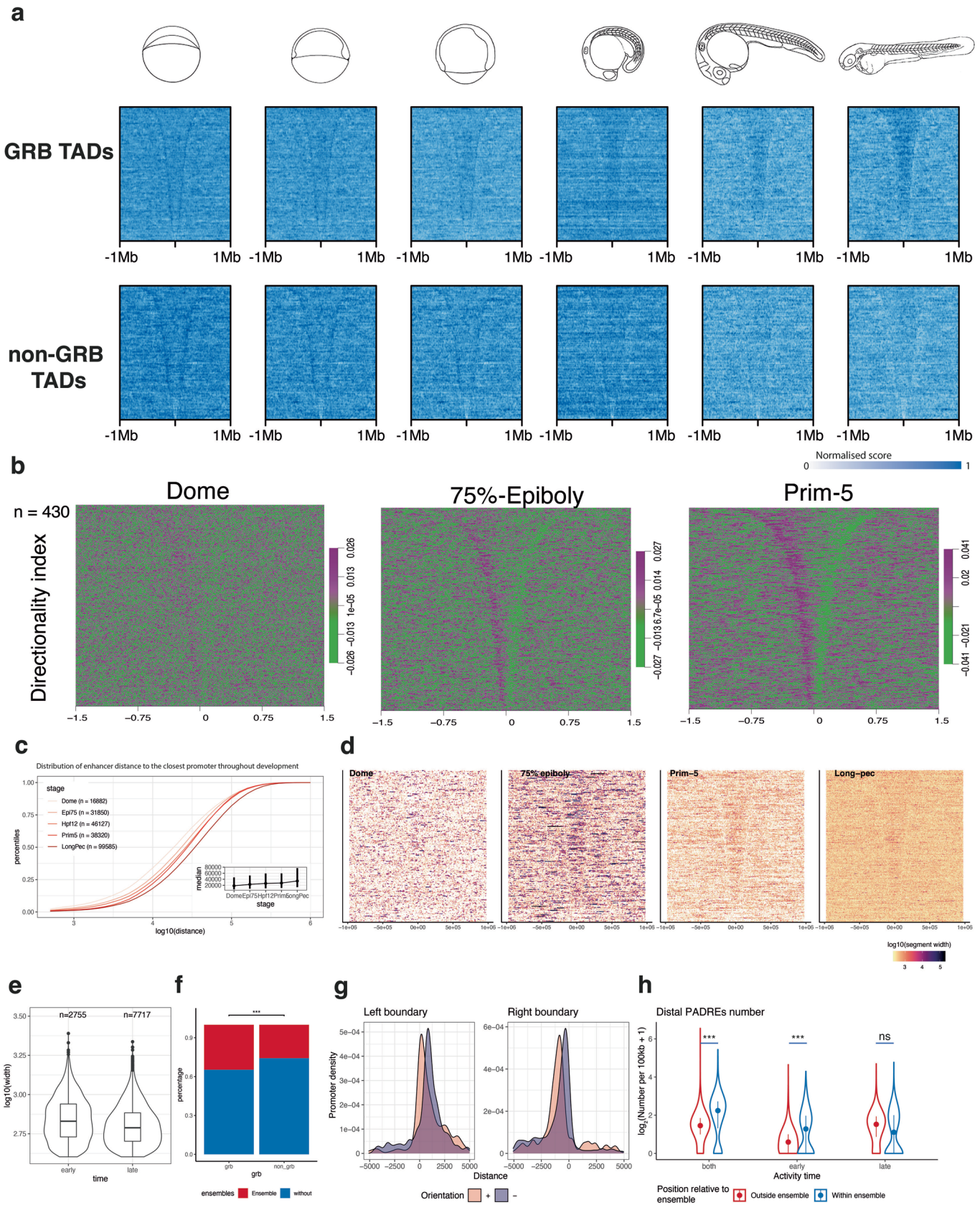


Extended Data Fig. 6 | UMAP visualisation of regulatory elements. a-c, Schematic representation of UMAP visualization of PADREs (for details, see Methods). R1-13 represent bins used to make the model. **d**, UMAP plot of annotated PADREs for each developmental stage analysed.



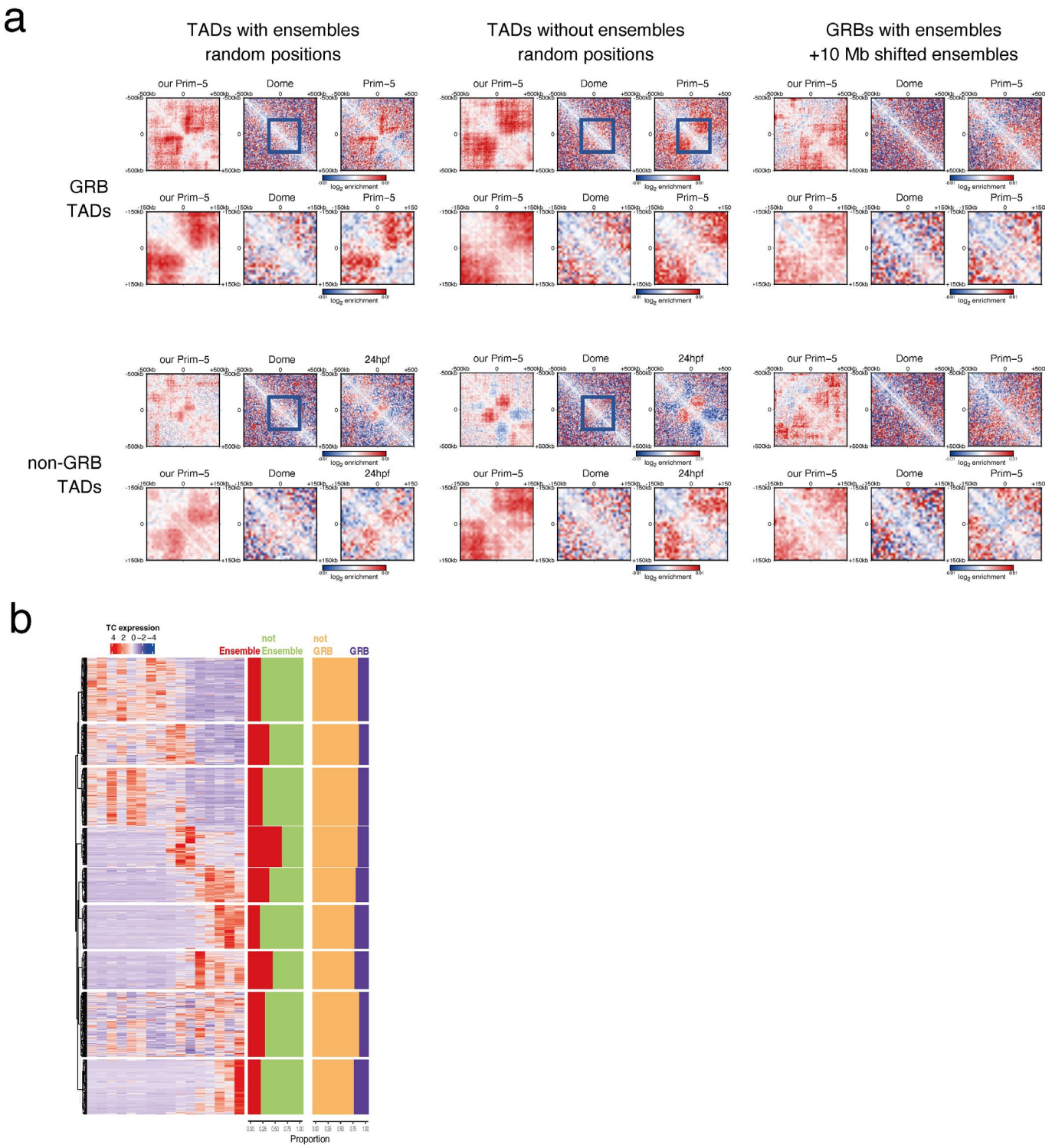
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Cell-type and developmental classification of PADREs. Cell-type specificity assignment and developmental dynamics of PADREs **a**, An example genomic region shows cell-type assignment of PADREs derived from single-cell ATAC-seq data (bottom track) and the gene model (top track). The name of PADREs contains their cell-type assignment. PADREs in this track are color-coded by their cell-type assignment as well, each colour representing a different cell-type. The state number in the name corresponds to those defined in McGarvey *et al.*[30]. **b**, Overlap of matches between the cell-type assignment and activity tissue determined by transgenic assay. Of 155 transgenically validated enhancers active at the Prim-5 stage, 117 have a cell-type specificity assignment. For details of anatomical terms and statistics see Supplementary Table 11. In 72 (62%) assigned transgenic enhancers the scATAC-seq derived anatomical annotation matches at least one of the activity domains of the transgenic reporter (left-hand side of the bar chart). Partial overlap indicates transgene activity in a related tissue, but without no identifiable direct overlap with that of the cell type assignment. Not assigned elements were not registered for cell type specificity by McGarvey *et al*.[30]. Undetermined elements were not possible to directly compare due to ambiguity of anatomical terms. The functional annotation of transgenically validated PADREs (right column) shows that most transgenic elements have an enhancer relevant ChromHMM registration at the Prim-5 stage. Waterfall plot between the left and right columns indicate overlap between cell type assignment and cis regulatory element category. **c**, Openness of distal (non-promoter) cPADREs throughout development at stages indicated on the x-axis in the defined SOM classes. Numbers in brackets indicate the number of elements in each class.

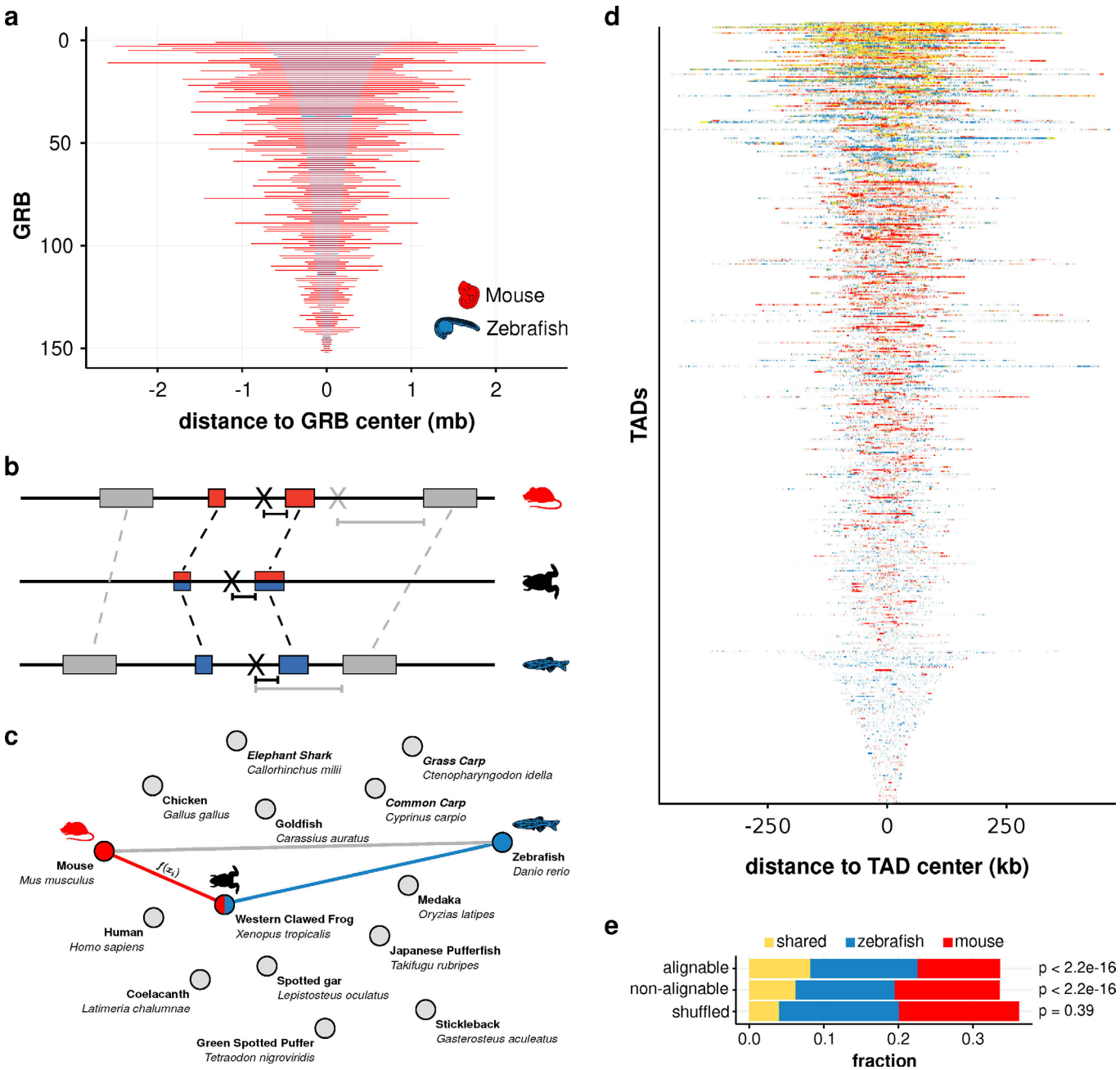


Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | Developmental dynamics of topologically associated domains and H3K27ac ensemble definition. a, ATAC-seq signals in GRB (top) and non-GRB (bottom) TADs throughout development. TADs are ordered in a descending order from the top of the heatmap. **b**, Directionality index in GRB TADs throughout development. **c**, Distance distribution of enhancer-associated PADREs to the closest promoter within GRB TADs. Bars represents inter-quartile range. **d**, Enhancer-associated ChromHMM segments in GRB TADs throughout development. TADs are ordered in a descending order from the top of the heatmap. Segments are coloured based on the logarithm of their length. Early stages are dominated by fewer large blocks, which start to be enriched within TADs only at 75%-Epiboly. In late stages, short segments are distributed uniformly throughout the entire TAD length. **e**, Width distribution of concatenated enhancer-associated ChromHMM segments. Singletons shorter than two bins (400 bp) were excluded. The number of segments is shown above each violin plot. **f**, Ratio of GRB and non-GRB tads containing H3K27ac ensembles. **g**, the density of CAGE promoters on ensemble boundaries. **h**, the number of non-promoter PADREs per 100 kb in TADs containing ensembles. The x-axis shows the developmental stage in which the PADRE is H3K27ac marked (early, late, or both). The location of promoters in respect to the ensemble is shown in different colours. The numbers were compared using two-sided two-sample unpaired Wilcoxon test.



Extended Data Fig. 9 | H3K27ac ensemble contact enrichment and CAGE expression patterns of gene classes separated by SOM. a, Controls for contact enrichment around H3K27ac ensembles. All regions were downsampled to n=56 to match the number of 50 kb - 150 kb size ensembles. Labels are as in Fig. 7g. The controls included random positions within the same TAD (**a**), random positions within TADs without ensembles, and 10MB shifted positions, for GRB TADs (top row) and non-GRB TADs (bottom row). The controls include published data for the Prim-5 stage, as well as new, unpublished data with higher resolution (Prim 5). **b**, CAGE expression patterns of gene classes separated by SOM. Bar plots in the middle show the proportion of ensemble-associated and GRB genes in each class respectively.



Extended Data Fig. 10 | Epigenetic domains comparison between zebrafish and mouse. a, Comparison of sizes of genomic sequences covering orthologous GRB-containing TADs. TADs are ranked by size, largest on top. **b**, Schematic illustration of the projection of an example genomic location X between zebrafish and mouse by interpolation using the direct alignments (grey rectangles) and the alignments via a bridging species (blue and red rectangles, Xenopus in this example). projections are indicated as a black X in the respective species). Dashed lines connect pairwise sequence alignments. The projected locations of X in mouse are indicated in grey (direct alignments) and black (via bridging species). **c**, Example graph comprising 15 species (nodes). For any genomic location, the shortest path through the species graph yields the combination of species which maximizes projection accuracy. **d**, H3K27me3 overlap profiles of all GRB TADs. TADs are ordered by their relative amount of shared signal. Bins are in the original genomic order. **e**, Fractions of bins with shared or species-specific H3K27me3 enrichment. Bins are classified as alignable (n=22,403) if they overlap a direct sequence alignment between zebrafish and mouse and as non-alignable otherwise (n=97,767). P-values are obtained by Fisher's exact test.

nature portfolio

Corresponding author(s): Dr. Ferenc Mueller
Dr. Boris Lenhard
Dr. Carsten O. Daub

Last updated by author(s): 2022-13-04

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

Data collection: DANIO-CODE DCC is based on *-DCC (https://gitlab.com/danio-code/public/dcc).
Data processing: The raw sequencing files were mapped using the DANIO-CODE pipelines (https://gitlab.com/danio-code,v1.1).
The genome track visualisations are based on WahsU Epigenome Browser and the UCSC genome browser. Further visualisation of chromatin marks and gene tracks was done using Gviz (https://github.com/ivanek/Gviz,v1.38.4).
CAGE data were further processed with CAGEr (https://bioconductor.org/packages/CAGEr/, v1.20). The enhancer were called using CageFightR (https://www.bioconductor.org/packages/CAGEfightR/,v1.8.0)
The 4C tracks were mapped using bowtie1 (https://github.com/BenLangmead/bowtie,v1.2.3).
The BS-seq were processed using Trimmomatic (https://github.com/usadellab/Trimmomatic,v0.38), sambamba (https://github.com/biod/sambamba,v0.6.8), samtools (https://github.com/samtools/samtools,v1.9), picard (https://github.com/broadinstitute/picard,v2.18.25), MethylDackel (https://github.com/dpryan79/MethylDackel,v0.3.0), UCSC kentUtils (https://github.com/ENCODE-DCC/kentUtils,v302.1 kentUtils 2014-11-11), deepTools (https://github.com/deeptools/deepTools,v3.1.1), and BEDTools (https://github.com/arq5x/bedtools2,v2.27.1)
RNA-Seq and aligned to GRCz10 using STAR aligner (https://github.com/alexdobin/STAR, v. 2.5.1b91).
StringTie was used to call transcripts (https://ccb.jhu.edu/software/stringtie/, v.1.33b92)
RNA-seq transcript were assembled using TACO (https://github.com/tacorna/taco,v0.7.3) and quantified using Salmon (https://github.com/COMBINE-lab/salmon, v0.11.2).
The genomic regions were segmented and annotated with ChromHMM (https://github.com/jernst98/ChromHMM, v1.12).
The nucleosome positions were calculated by NucleoATAC (https://github.com/GreenleafLab/NucleoATAC, v0.3.4).
CTCF binding sites were identified using TFBSTools (https://www.bioconductor.org/packages/TFBSTools/, v1.28)
danRer10 to danRer 11 lift over was performed using the rtracklayer package (https://bioconductor.org/packages/rtracklayer/,v1.54.0)
Self-organizing-maps were generated using the kohonen package (https://cran.r-project.org/web/packages/kohonen,v.3.0.8)
The genomation package was used to calculate the mean nucleosome signal of the cPADRES as well as to visualise the ATAC and ChIP-seq signals (https://bioconductor.org/packages/genomation/,v1.26.0) and visualised using the uwot (https://cran.r-project.org/web/packages/uwot/, v0.1.10).

Peak summits were defined using MACS2 (https://github.com/macs3-project/MACS, v2.2.7.1).
The Hi-C data was processed using HiCUP (https://github.com/StevenWingett/HiCUP,v0.6.0) and mapped using Bowtie2 (https://github.com/BenLangmead/bowtie2,v2.3.4.1). The aligned data were further processed using HOMER (http://homer.ucsd.edu/homer,v4.11).
Hi-C matrices were processed using FAN-C (https://github.com/vaquerizaslab/fanc,v0.9.0). TADs were called using TADtool (https://github.com/vaquerizaslab/tadtool, v.0.76). The heatmap for CAGE promoters in TADs were generated by the ComplexHeatmap package (https://www.bioconductor.org/packages/ComplexHeatmap/, v0.75).
The Enhancer ensembles were detected using the ROSE algorithm (https://bitbucket.org/young_computation/rose, commit feb35cb1d9556a76f8ac1f51521539bb30651343). The H3K27ac signal across TADs was visualised using the heatmaps package (https://bioconductor.org/packages/heatmaps/, v1.18.0)
Motif hits for the conservation analysis were computed using seqPattern (https://bioconductor.org/packages/seqPattern, v1.26.0).

Data analysis

The code used for data analysis is available in the following GitHub repository:
https://github.com/DANIO-CODE/DANIO-CODE_Data_analysis

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All the data are available on DANIO-CODE Data Coordination Center (https://danio-code.zfin.org). The previously unpublished data are available under the following IDs: DCD000519SR, DCD000518SR, DCD000514SR, DCD000493SR, DCD000490SR, DCD000244SR
The raw sequencing data produced for this study are available on the European Bioinformatics Institute (EBI) European Nucleotide Archive (ENA) under study numbers PRJNA824720, PRJNA821001, PRJNA821088, PRJNA821148, and PRJNA821034.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was determined by availability of genomics datasets which were collected from public databases.
Data exclusions	Genomics data were excluded which did not meet the quality control criteria (as set by ENCODE criteria) and which when compared by PCA was identified as technical outlier.
Replication	Genomics data were used where at least 2 replicates were available. The complete datasets consists of two parts: 1. publicly available datasets collected and re-processed in a standardised way as a part of this study. The replication in those cases were constrained to data availability. All data used in the analysis of this paper passed the quality control. 2. Newly generated data for this study: All of the newly generated data for this study passed the quality control and reproducibility controls, except for 2 H3K4me3 Long-Pec samples, which subsequently haven't been used in further analyses. Summary of all datasets and their numbers is available in Figure 1. Details about all datasets, as well if they were used in the analysis for this paper are available as Supplementary Table 1.
Randomization	Not applicable due to the nature of analyses of genomics datasets (there was no case/control experimental design in this study). The zebrafish embryos were collected by standard laboratory procedure described in the Supplementary methods section.
Blinding	Not applicable due to the nature of analyses of genomics datasets (there was no case/control experimental design in this study). The zebrafish embryos were collected by standard laboratory procedure described in the Supplementary methods section..

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☐	☒ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used

Polyclonal rabbit anti-H3K4me3 from Diagenode, Cat-No. C15410003
Polyclonal rabbit anti- H3K27ac from abcam, Cat-No. ab4729
Polyclonal rabbit anti-H3K4me1 from abcam, Cat-No. ab8895
Polyclonal rabbit anti-H3K27me3 from Diagenode, Cat-No. C15410069
Polyclonal rabbit anti-H3K36me3 from abcam, Cat-No. ab9050
Polyclonal rabbit anti- Pol II (phospho S5) from abcam, Cat-No. ab5131
Polyclonal rabbit anti-H2A.Z from abcam, Cat-No. ab4174

Validation

Polyclonal rabbit anti-H3K4me3 from Diagenode, Cat-No. C15410003 was validated for ChIP-qPCR by the manufacturer using human K562 cells. Species reactivity: Human, mouse, zebrafish, trout, Daphnia, Arabidopsis, rice, tomato, maize, poplar, silena latifolia.
Polyclonal rabbit anti- H3K27ac from abcam, Cat-No. ab4729 was validated for X-ChIP by the manufacturer using HeLa cells. Predicted reactivity with Rat, Chicken, Xenopus laevis, Arabidopsis thaliana, Drosophila melanogaster, Monkey, Zebrafish, Plasmodium falciparum, Rice, Cyanidioschyzon merolae.
Polyclonal rabbit anti-H3K4me1 from abcam, Cat-No. ab8895 was validated for X-ChIP by the manufacturer using U-2 OS cells and previously reported to cross-react with zebrafish samples.
Polyclonal rabbit anti-H3K27me3 from Diagenode, Cat-No. C15410069 was validated for ChIP by the manufacturer using K562 cells. Species reactivity: Human, mouse, rat, pig, zebrafish, Drosophila, Schistosoma, Arabidopsis, cow
Polyclonal rabbit anti-H3K36me3 from abcam, Cat-No. ab9050 was validated for X-ChIP by the manufacturer using U-2 OS cells. Predicted reactivity: Mouse, Rat, Saccharomyces cerevisiae, Xenopus laevis, Arabidopsis thaliana, Caenorhabditis elegans, Drosophila melanogaster, Plants, Schizosaccharomyces pombe, Zebrafish, Silk worm, Rice, Xenopus tropicalis, Trypanosoma brucei.
Polyclonal rabbit anti- Pol II (phosphor S5) from abcam, Cat-No. ab5131 was validated for X-ChIP by the manufacturer using U-2 OS cells. Predicted reactivity: Mouse, Rat, Saccharomyces cerevisiae, Xenopus laevis, Arabidopsis thaliana, Caenorhabditis elegans, Drosophila melanogaster, Schizosaccharomyces pombe, Zebrafish, a wide range of other species.
Polyclonal rabbit anti-H2A.Z from abcam, Cat-No. ab4174 was validated for X-ChIP by the manufacturer using HeLa cells. Predicted reactivity: Sheep, Xenopus laevis, Arabidopsis thaliana, Zebrafish.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals

In this study, 6-18 month old zebrafish males and females belonging to the Ab and Buc-GFP strains were exclusively used for natural breeding. Fish were maintained in a designated facility according to the UK Home Office regulations and UK Animals (Scientific Procedures) Act 1986. All animals were kept in a recirculating system (ZebTEC, Tecniplast) in a 10-hour dark, 14-hour light photoperiod at 26°C and fed 3 times daily.
Zebrafish embryos were obtained by crossing one male and one female (separated by a divider) in a specifically designed crossing cage in the evening. On the next morning, dividers were removed, and embryos produced by natural spawning were collected shortly (5-10 min) after laying by using a small net.

Wild animals

No wild animals were used in this study.

Field-collected samples

No field collected samples were used in this study.

Ethics oversight

Wild type (AB) and Tg(Buc-GFP) strains were maintained in a designated facility according to the UK Home Office regulations and UK Animals (Scientific Procedures) Act 1986. Fish were kept in a recirculating system (ZebTEC, Tecniplast) in a 10-hour dark, 14-hour light photoperiod at 26°C and fed 3 times daily.
Zebrafish experiments were restricted to embryonic stages and adults were only used for natural breeding under the project licences assigned to the universities involved in this project. Zebrafish embryos were obtained by crossing one male and one female (separated by a divider) in a specifically designed crossing cage in the evening. On the next morning, dividers were removed, and embryos produced by natural spawning were collected shortly (5-10 min) after laying by using a small net.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	https://danio-code.zfin.org/dataExport/daniocode/detailSeries/518/
Files in database submission	Along with raw fastq reads, for each sample a signal (p-value of signal over control) file in bigWig format is provided, along with peaks in narrowPeak file format.
Genome browser session (e.g. UCSC)	Processed data can be viewed in the DANIO-CODE Track Hub, listed in the UCSC Genome browser public Track Hubs, or uploadable through the link: https://danio-code.zfin.org/trackhub/DANIO-CODE.hub.txt

Methodology

Replicates

5 - 9 somites samples and Long-Pec samples (H3K27ac, H3K4me1, and H3K27me3) were done in 1 replicate, as these samples were already available in 1 replicate from before.
Long-pec H2AZ, PolII, H3K4me3 were done in 2 replicates.

Sequencing depth

All samples were 75bp paired-end datasets.
The sequencing depth for each sample after optical duplicates removal:

5-9 somites	H3K27me3	26938306
	H3K4me1	14905594
	H3K36ne3	43630140
Long-pec H2AZ	rep1	18712668
	rep2	36137656
	H3K4me3 rep1	13346874
	rep2	21845236
	PolII rep1	41158384
	rep2	16735286
	H3K27ac	58048920
	H3K27me3	59785820
	H3K4me1	53491012

Antibodies

Polyclonal rabbit anti-H3K4me3 from Diagenode, Cat-No. C15410003
Polyclonal rabbit anti- H3K27ac from abcam, Cat-No. ab4729
Polyclonal rabbit anti-H3K4me1 from abcam, Cat-No. ab8895
Polyclonal rabbit anti-H3K27me3 from Diagenode, Cat-No. C15410069
Polyclonal rabbit anti-H3K36me3 from abcam, Cat-No. ab9050
Polyclonal rabbit anti- Pol II (phospho S5) from abcam, Cat-No. ab5131
Polyclonal rabbit anti-H2A.Z from abcam, Cat-No. ab4174

Peak calling parameters

Program: macs2 callpeak
Parameters: -f BED -g 1371719383 -p 0.01 --nomodel --shift 0 --extsize 205 --keep-dup all -B --SPMR

Data quality

For the following samples fraction of reads in peaks was used as a measure of quality:

5-9 somites	H3K27me3	0.2323
	H3K4me1	0.3395
	H3K36ne3	0.2524
Long-pec H2AZ	rep1	0.2539
	rep2	0.2539
	H3K4me3 rep1	0.0071 - not used in analysis
	rep2	0.0071 - not used in analysis
	PolII rep1	0.0256
	rep2	0.0256

For the following samples, the proportion of peaks overlapping in pseudo-replicates was used as a method of quality check:

Long-pec	H3K27ac	0.5935
	H3K27me3	0.2562
	H3K4me1	0.5098

Software

The code and detail for ChIP-seq processing are available on: https://gitlab.com/danio-code/DANIO-CODE_ChIP-seq