From Understudied to Understood: Multi-Omics Analysis with MiniENCODE Exemplified by Zebrafish

Hang Yang[1,2], Zhaojun Shan[1,2], Hanqiao Shang[1], Puxuan Jiang[1,2], Yingshu Li[1,*], Qiang Tu[1,2,*]

[1] *State Key Laboratory of Molecular Developmental Biology, Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China* [2] *University of Chinese Academy of Sciences, Beijing 100049, China* [*] *Corresponding authors: ysli@genetics.ac.cn, qtu@genetics.ac.cn*

Abstract

The ENCODE project provides a comprehensive map of genomic activity in humans and mice. However, its extensive toolbox can be too costly and technically challenging for smaller research communities. This presents a hurdle for understudied model organisms where resource constraints and complex omics data integration further compound the challenge. To bridge this gap, our study introduces a miniENCODE project, using the zebrafish as an exemplar. We collected, generated and integrated datasets from three experimental assays across developmental stages and adult tissues. This analysis is facilitated by our newly developed miniODP database, a computational tool designed for the integration and visualization of multi-omics data. Utilizing these methods, we have cataloged over 52,000 candidate enhancers, identified various stage-specific, tissue-specific, and constantly active enhancers, and experimentally validated a subset of heart-specific enhancers. We have delineated key transcription factors and their corresponding regulatory networks across developmental stages and adult tissues. This study not only delivers valuable regulatory insights for the zebrafish research community but also illustrates how an integrative approach can aid in understanding complex regulatory mechanisms even with limited resources. This strategy could empower scientific communities working with various understudied model organisms to expedite their genomic research without overstretching their resources.

Introduction

The Encyclopedia of DNA Elements (ENCODE) project, a public research consortium, was initiated after completing the Human Genome Project in 2003. It aims to comprehensively identify all functional elements in the human genome [1]. Phase III of the ENCODE project has yielded a vast amount of omics data, and developed a registry of 926,535 human and 339,815 mouse candidate *cis*-regulatory elements (CREs). Furthermore, a comprehensive database was constructed to facilitate access to the registries [2]. Analogous efforts have also been devoted to analyzing the genomes of *C. elegans* and *D. melanogaster* through the modENCODE Project [3, 4]. Such resources are of great value to the scientific communities that employ these model organisms, with a significant impact on their research.

A broad array of other model organisms, beyond those currently represented in the ENCODE and modENCODE projects, is employed in research. However, their omics data remain relatively limited. As evidenced by the GEO database, approximately 80% of the samples are derived from humans and mice. Only 9% originate from the next eight most represented organisms, while the remaining 11% are from more than 6,000 organisms (Supplementary Figure 1). Consequently, communities working with less commonly studied organisms face significant challenges in obtaining comprehensive information on genome activities, given their limited research resources. Additionally, experimental biologists confront a daunting task when exploring the published omics datasets of the model organisms they work with.

Consider the zebrafish (*Danio rerio*) as an example: this powerful model organism used is utilized in various fields, including developmental biology, toxicology, human diseases, regeneration, drug discovery, and many others [5]. A plethora of omics datasets has been generated, as highlighted in a recent study [6]. A variety of genomics and functional genomics databases for zebrafish have been developed. For instance, the Zebrafish Information Network (ZFIN) provides a comprehensive collection of expertly curated and cross-referenced zebrafish genetic and genomic data, covering genes, alleles, transgenic lines, gene expression, gene function, pheno-

types, orthology, human disease models, nomenclature and reagents [7]. The DANIO-CODE consortium creates a track hub to consolidate multiple zebrafish genomic and epigenomic datasets [8, 9]. The Expression Atlas [10] contains information on gene expression patterns, including microarray and RNA-seq data under different biological conditions and species, including only 25 datasets from zebrafish. Single Cell Expression Atlas [11] contains reanalyzed and annotated single-cell RNA-Seq data across multiple species, including 14 zebrafish datasets totaling 142,000 cells. Although these databases are invaluable resources for researchers in the zebrafish community and beyond, they primarily focus on collecting general zebrafish genomic and genetic information, or multi-omics visualization through tracks. Integrated analysis and visualization tools are notably absent.

The key to solving these challenges is utilizing existing resources more efficiently. This objective necessitates two critical components: firstly, the generation of pivotal and informative omics data via a selective set of technologies; secondly, the facilitation of reuse and sharing of published omics data. In response to these challenges, we have proposed two solutions: miniENCODE experimental assays and miniODP (mini Omics Data Portal) database. The miniENCODE experimental assays combine state-of-the-art omics technologies, many of which were employed and refined by the ENCODE project. Meanwhile, the miniODP database integrates computational tools to collect, reanalyze, and visualize published omics datasets. Together, these two solutions empower us to conduct a miniENCODE project for any model organism of interest with limited resources.

Consequently, we performed a miniENCODE analysis of zebrafish in the current study. We compiled both published and our datasets from three miniENCODE experimental assays (RNA-seq, ATAC-seq, and ChIP-seq of H3K27ac), covering six developmental stages and 11 adult tissues. We then developed miniODP, a computational platform facilitating multi-omics datasets' integration, analysis, and visualization. Through this analysis, we compiled a registry of 52,350 candidate enhancer-like signatures (ELSs), and elucidated major regulators and gene regulatory networks (GRNs) activated during these developmental stages and in the tissues. Supported by the experimental validation and corroborated by the existing literature, our findings illustrate that a miniENCODE analysis provides a pragmatic and efficient approach for identifying candidate enhancers and exploring regulatory mechanisms in a given organism.

Results

MiniENCODE assays

The ENCODE project has harnessed an array of technologies to reveal functional DNA elements in genomes, including those involving DNA-protein interactions, transcription, chromatin accessibility, DNA methylation, RNA binding, 3D chromatin structure, and replication timing [2]. However, for small research communities, deploying all these technologies is neither feasible nor necessary. Furthermore, determining the essential technologies for initial profiling can be perplexing for biologists with limited omics expertise.

To address this issue, we propose using a select set of omics assays as an initial choice for models lacking data. These technologies are straightforward to implement, do not demand costly reagents (e.g., high-quality antibodies), and complement one another to offer comprehensive information across different facets of genome activities. We refer to this set of omics profiling methods as miniENCODE assays (Figure 1A), which include: (1) Short-read RNA-seq for quantifying gene expression; (2) ATAC-seq for identifying potential regulatory elements; (3) CUT&Tag of specific histone modifications (H3K4me3, H3K27ac) to annotate the types of regulatory elements; (4) Long-read RNA-seq for constructing accurate gene models (optional); (5) Micro-C for uncovering 3D interactions among regulatory elements (optional); and (6) Single-cell RNA-seq, single-cell ATAC-seq, single-cell CUT&Tag for cell-type specific profiling (optional).

Short-read RNA-seq, ATAC-seq, and ChIP-seq are fundamental technologies that yield rich information across diverse aspects of genome activities. They are widely utilized in various organisms. Notably, CUT&Tag, a recent improvement over traditional ChIP protocols, excels in resolution, signal-to-noise ratio, sequencing depth requirement, and workflow simplicity [12]. For the identification of regulatory elements, the most informative types of histone modification are H3K4me3 and H3K27ac, which were employed in CRE prediction by the ENCODE project [2].

In addition to the essential miniENCODE assays, we also proposed three optional assays. Despite being relatively expensive and complex, these assays can offer important insights depending on the research objectives. Long-read RNA-seq is a crucial technique for analyzing newly sequenced genomes. Constructing accurate gene models is critical for many downstream genomic analyses. Long-read RNA-seq technologies yield single reads covering the entire mRNA molecules, thus circumventing the error-prone computational assembly used in short-read RNA-seq gene model building [13]. Micro-C is a recently improved method that outperforms traditional Hi-C technologies regarding resolution and sequence depth required

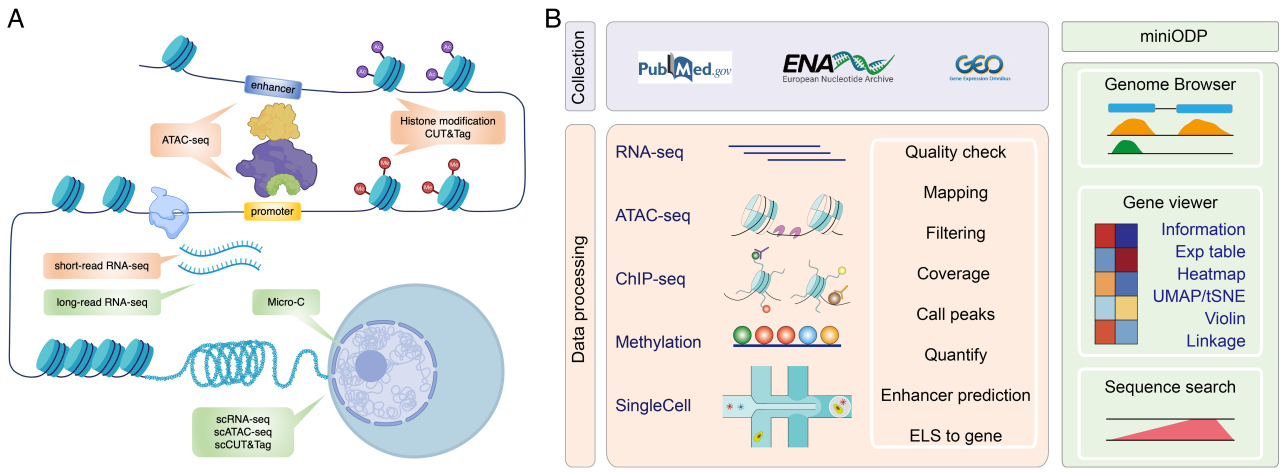

Figure 1: *Overview of miniENCODE assays and the miniODP database. (A) Schema illustrating the selected technologies employed in the miniENCODE assays. Essential technologies are depicted in salmon-colored boxes, and optional technologies in green-colored boxes. This graphics was created in part using BioRender.com. (B) Layout of the miniODP database. The database consolidates multi-omics data from public resources, processes it using a uniform computational pipeline, and generates derived datasets, such as ELS prediction. The miniODP provides a user-friendly web interface for these datasets' visualization.*

[14, 15]. This technology can be deployed to investigate long-range interactions between enhancers and promoters. Lastly, single-cell RNA-seq, ATAC-seq, CUT&Tag technologies deliver information on gene expression, chromatin accessibility, and histone modification at the single-cell level, which is crucial to understand the regulatory mechanisms in complex tissues [16–19].

MiniODP database

Omics datasets have been generated at an unprecedented speed. Scientific communities working with a few major model organisms, such as humans or mice, have many specialized secondary databases. However, the reuse, sharing, and integration of data for comprehensive analysis remain challenging for communities using less common model organisms. Even for many model organisms with central portal websites, e.g., zebrafish and medaka, databases for omics data reuse and integrated analysis are still scarce. In response, we developed miniODP, a generalized database that includes a suite of computational tools for analyzing, integrating, and visualizing multi-omics data (Figure 1B, `https://tulab.genetics.ac.cn/miniodp/`).

MiniODP offers three modules for user access to different data types. The first module, a JBrowse-based genome browser [20], ensures a responsive interface with a sophisticated track selector to visualize omics data tracks of various technologies and samples (Figure 2A-B). Users can swiftly navigate to a locus of interest via gene name or genomic coordinate searches, and use the track selector to filter hundreds of tracks based on category, investigator, sample, source, target, technique, and type.

The second module is a gene browser that delivers detailed information about given genes, including functional annotation, quantitative expression tables, and plots. This module, developed using R Shiny and ShinyProxy for the front-end and an SQLite database as the backend, allows searching for genes by Ensembl ID, name, GO annotation, Interpro protein domain annotation, transcription factor (TF) family and orthologs (Figure 2C). The results link to the genome browser, Ensembl [21], AnimalTFDB [22] and InterPro [23] databases. Quantitative expression data associated with selected genes can be visualized as heatmaps or line plots (Figure 2D). Furthermore, single-cell datasets can be represented as UMAP, t-SNE, violin plots, and heatmaps (Figure 2E). Finally, gene expression, chromatin accessibility, and ELS to gene linkages around the given locus can be visualized as a detailed multi-omics plot (Figure 2F).

The third module is a sequence search tool, implemented using SequenceServer [24]. This module enables users to search sequences in genomes, cDNAs, and reference proteomes in the database (Figure 2G). Hits in results are linked to the genome browser, aiding the interpretation of the results in their genomic context (Figure 2H).

MiniENCODE project for zebrafish

We carried out a miniENCODE project for zebrafish, applying the miniENCODE experimental assays and utilizing our developed miniODP database. As previously discussed, despite the abundance of omics datasets generated for zebrafish by the scientific community, a data portal for the presentation and integration of these datasets is currently lacking. Therefore, our miniENCODE analysis on zebrafish is a testa-

Figure 2: *Modules of miniODP database. (A, B) Genome browser, equipped with a sophisticated track selector, allows visualization of diverse omics data tracks. (C-F) Gene browser for querying gene-specific annotations and visualizing gene expression patterns, including single-cell RNA-seq and multi-omics data. (G, H) BLAST server for sequence search. BLAST results link back to the genome browser, which enables the user to interpret BLAST search results in the genomic context. Integrating these modules enhances the user's ability to extract biological insights from the vast amount of omics data.*

ment to the effectiveness of a limited number of omics techniques coupled with an integrated database for discerning the regulatory elements encoded within a genome.

Zebrafish omics datasets were collected from various public databases, such as Pubmed, GEO [25], and ENA [26]. Each dataset was manually curated to assess the meta-information including the original publication, experimental design, and biological samples. Our focus centered on several data types, namely RNA-seq, ATAC-seq, ChIP-seq, DNA methylation, scRNA-seq, and scATAC-seq data, which were selected based on criteria like data quality, the technology utilized, biological samples profiled, and the number of time points surveyed. We then established a data type matrix of biosamples and technologies, illustrating the datasets' distribution (Figure 3A). Samples were categorized into broader classes, including eight developmental stages, 20 embryonic and adult tissues, and others. Technologies, particularly ChIP-seq in-

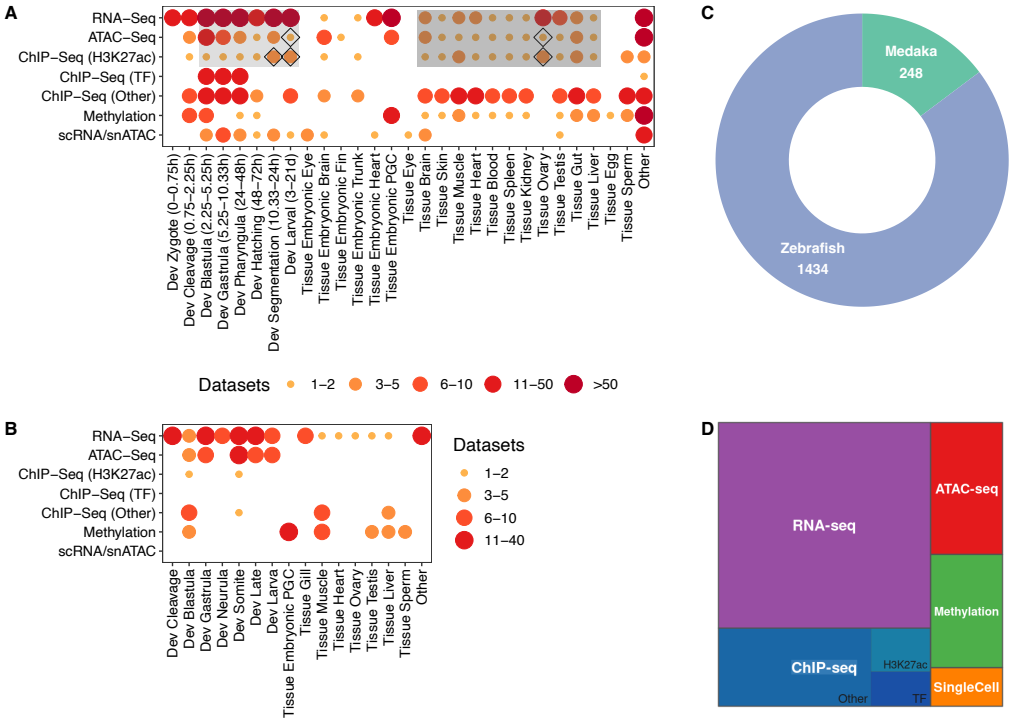

Figure 3: *Summary of miniODP modules for zebrafish and medaka. (A) Data type matrix of biosample and technologies categories for zebrafish. Dot size and color represent the number of datasets of the given biosample and technology category. Light and dark grey boxes denote essential miniENCODE assays for the studied developmental stages and adult tissues, respectively. Diamonds mark datasets generated in this study. (B) Data type matrix for medaka. (C-D) Summaries of data by species and technology.*

cluding CUT&Tag, were categorized based on assay targets: histone modification H3K27ac, TFs, etc. Our primary objective was to cover three essential miniENCODE assays for six major developmental stages (light grey box in Figure 3A) and 11 major adult tissues (dark grey box in Figure 3A). To fill the few missing datasets, we generated ATAC-seq and H3K27ac CUT&Tag datasets for the segmentation stage, larval stage, and ovary tissue (diamonds in Figure 3A, Supplementary Figure 2 and Supplementary Figure 3), thereby completing our primary objective matrix (3 assays $\times$ 17 biosamples). In addition, we also collated multi-omics data for medaka, another widely used fish model organism (Figure 3B).

Altogether, we obtained 1,624 bulk datasets comprising 233 billion reads and 58 single-cell RNA-seq or ATAC-seq datasets encompassing 385,000 cells. These datasets encompassed 1,434 for zebrafish and 248 for medaka (Figure 3C). The collection, regarding technology, comprised 910 RNA-seq, 346 ChIP-seq, 198 ATAC-seq, 170 methylations, and 58 single-cell datasets (Figure 3D). In total, 858 tracks presenting original and derived datasets were generated, with 728 for zebrafish and 130 for medaka.

The vast multi-omic data collected facilitated an integrated analysis. Firstly, ELSs were defined using ATAC-seq and H3K27ac ChIP-seq data. Linkages between ELSs and target genes were then predicted by integrating RNA-seq data. Further integrating TF binding motifs, we identified the major regulators and regulatory networks for each stage and tissue.

Exploring enhancer-like signatures

Enhancers play a crucial role in orchestrating the spatiotemporal expression of genes that underlie biological processes. They are typically associated with open chromatin regions marked by specific histone modifications and are bound by cell type-specific TFs [27]. Such characteristics permit genome-wide identification of enhancers. ATAC-seq, a technique that uses hyperactive Tn5 transposase to insert sequencing adapters into accessible chromatin regions, allows for the detection of enhancer regions due to their hypersensitivity to nucleases [28]. Active enhancers are typically associated with histone modifications, primarily H3K27ac, which can be detected via techniques like ChIP-seq or CUT&Tag [12, 29].

Leveraging the collection of omics datasets that encompass three essential miniENCODE assays across six stages and 11 tissues, we predicted candidate enhancers using the computational method derived from the ENCODE project [2]. Initially, we pinpointed representative DNA accessibility regions (rDARs) utilizing ATAC-seq peaks. For each rDAR, we calculated the Z-scores for the log10-transformed ATAC-seq and

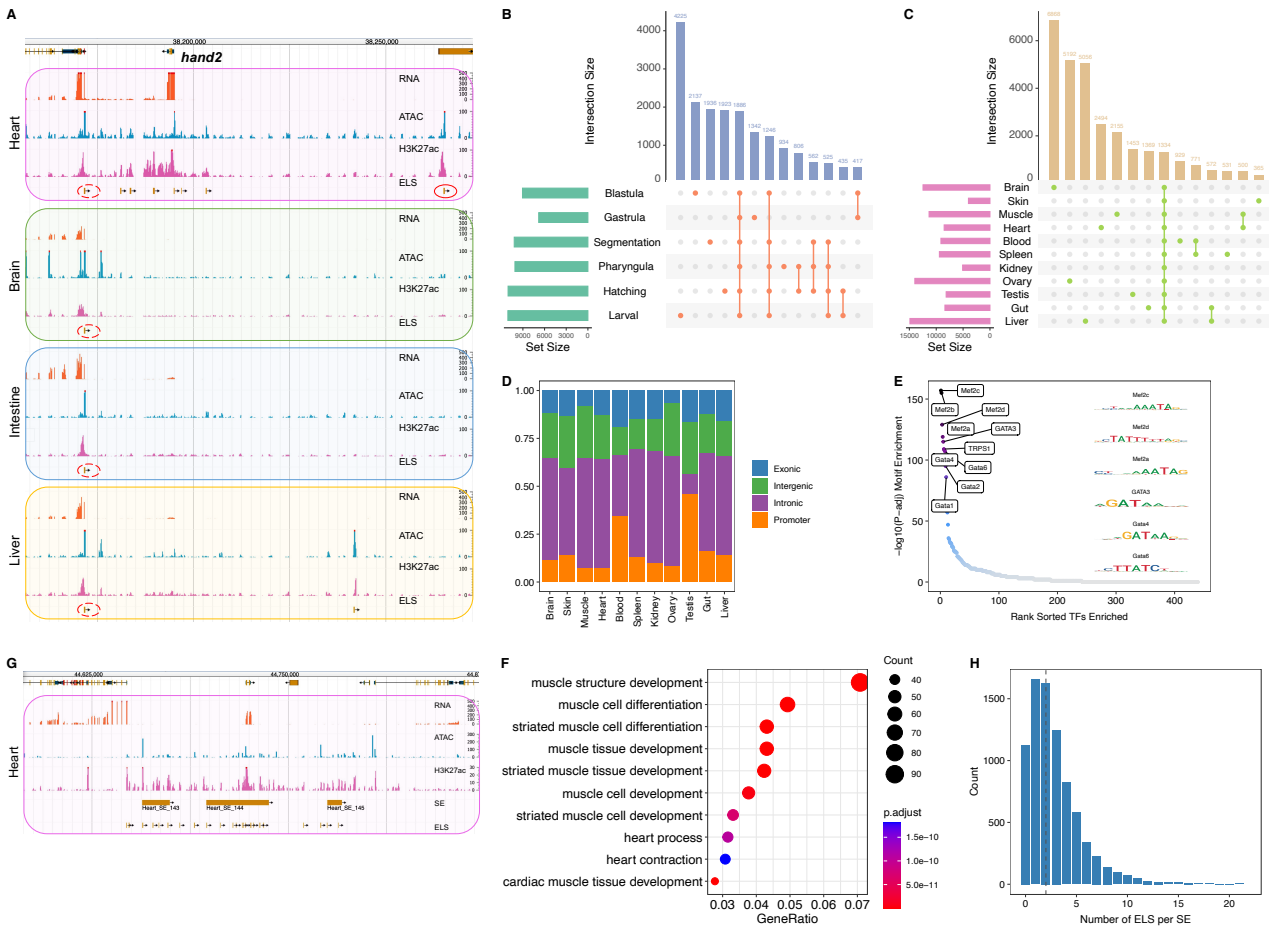

Figure 4: *Enhancer-like signatures of zebrafish. (A) Genomic view of ELSs across tissues supported by multi-omics data. Seventeen tracks are displayed, beginning with the gene annotation track at the top. Subsequent panels are dedicated to four tissue-specific profiles: heart, brain, intestine, and liver. Each panel presents four omics data tracks: RNA-seq (coral peaks), ATAC-seq (sky blue peaks), ChIP-seq for H3K27ac (light pink), and predicted ELSs (mustard yellow blocks). The identification of ELSs is substantiated by the presence of both ATAC-seq and H3K27ac signals. Notably, on the left is a ubiquitously active ELS (dashed ellipse) present across all four tissues, contrasted on the right by a heart-specific ELS (solid ellipse) which has been experimentally validated in this study. (B-C) Upset plots demonstrate the intersections of numbers of ELSs activated in different developmental stages (B) or tissues (C). Horizontal bars represent individual stages or tissues, with their length (set size) indicating the number of ELSs activated in the given stage or tissue. Intersections between sets are shown by connected dots, with vertical bars (intersection size) indicating the number of ELSs in the intersection. During development, the larval stage was notable for the highest counts of stage-specific ELSs, whereas the brain tissue exhibited the most tissue-specific ELSs. Over 1,000 ELSs were consistently activated across all the analyzed stages or tissues. (D) Genomic distribution of ELSs. The majority of tissue-specific ELSs were predominantly situated within intronic (51%) or intergenic regions (23%). (E) Enrichment analysis of TF binding motifs within heart-specific ELSs. TF motifs are ranked by the significance of their enrichment. The top 10 TFs are labeled, and the motifs of the six TFs known to be functional in the heart are presented. (F) Gene Ontology analysis of genes linked to heart-specific ELSs. (G) The genomic view of multiple super-enhancers. Each consists several ELSs. These super-enhancers and their adjacent regions are characterized by strong ATAC-seq and H3K27ac signals, indicating regions of high regulatory activity. (H) Distribution of the number of ELSs per super-enhancer.*

H3K27ac signals in each biosample. Subsequently, we averaged the H3K27ac signals across an extended area that included the rDAR and a 500 bp margin on both sides. Enhancer-like signatures were consequently defined as rDARs with a Z-score threshold of 1.64 of both ATAC and H3K27ac signals, corresponding to the 95th percentile. ELSs were subsequently categorized as either TSS-proximal with enhancer-like signatures (pELS, 0-2000 bp from the transcription start site) or TSS-distal with enhancer-like signatures (dELS, >2000 bp from the transcription start site). This process culminated in creating a comprehensive register of 52,350 ELSs in the zebrafish genome, which represent 1.2% of the genome and encompass 17% of the ATAC-seq signal peaks. Out of these ELSs, 37% were designated as pELSs, whereas 63% were classified as dELSs. To the best of our knowledge, this represents the most comprehensive registry of ELSs in the zebrafish genome to date, covering both development and adult tissues. These ELSs are accessible via the

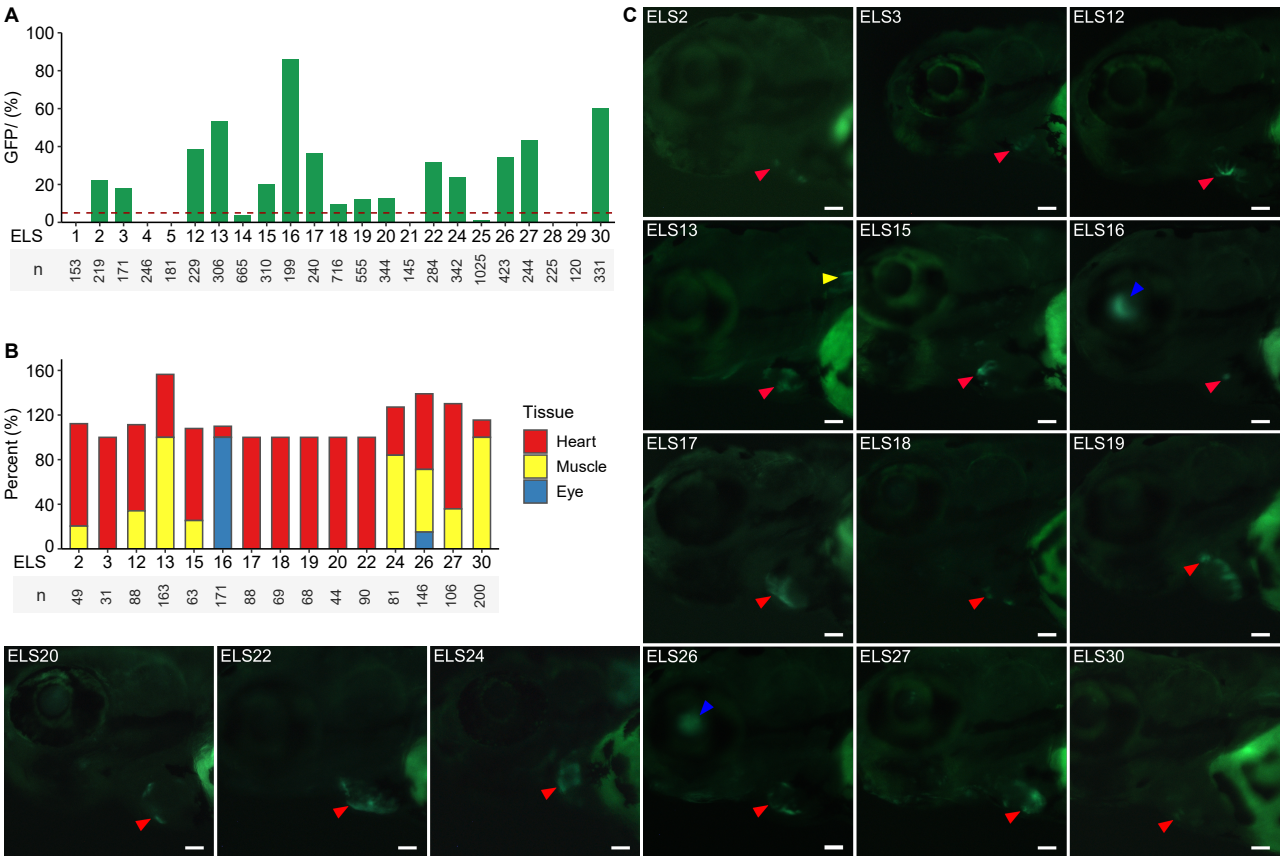

Figure 5: *Reporter assay validating in vivo function of heart-specific ELSs. (A) Percentage of GFP-positive embryos among total observed embryos. 15 out of 23 ELSs could drive GFP expression in over 5% (dashed line) embryos observed. (B) Percentage of GFP signals observed in different organs among GFP-positive embryos. GFP signals were predominately observed in the heart, with muscles and eye expression also detected. (C) Image of representative embryos for each ELS in report assays. Arrows indicate GFP signals matching the color code in (B). Scale bar,* 50 µm.

database's genome browser (Figure 4A).

Enhancers, bound by cell type-specific TFs, are instrumental in determining the precise timing and location of gene expression [30]. As such, many enhancers display cell-type-specific activities. We classified these ELSs based on the developmental stages or tissues in which they are active. Among the identified ELSs, we found 51% (26,695 ELSs) exhibiting stage- or tissue-specific activity. In (Figure 4A), we illustrate both a ubiquitously active ELS and a heart-specific ELS, each supported by robust ATAC-seq and H3K27ac signals. The heart-specifc ELS, referred to as ELS22 in this study, was experimentally validated (see below). This underscores the importance of integrating multi-omics data across multiple tissues to predict ELSs accurately. Notably, during development, the larval stage accounted for the highest number of stage-specific ELSs (Figure 4B), likely due to the cell type diversity and chromatin dynamics characteristic of this late developmental stage. Within adult tissues, the brain, ovary, and liver contained the highest number of tissue-specific ELSs (Figure 4C), suggesting increased heterogeneity and active regulatory processes in these tissues.

Intriguingly, we observed a considerable overlap of ELSs across developmental stages (evidenced by multiple vertical lines connecting dots in Figure 4B). At the same time, there was minimal overlap between ELSs in distinct tissues (few vertical lines connecting dots in Figure 4C). Temporally contiguous stages of development feature a higher degree of similarity in their gene expression profiles, reflecting the need for coordinated gene regulation to drive proper cell differentiation and tissue development. In contrast, spatially discrete organs share fewer ELSs, likely reflecting their unique functional and structural attributes. Shared ELSs were found in organs with similar developmental origins and related functions: muscles and hearts, both mesoderm-derived and contractile; blood cells and spleen, both emerging from the lateral plate mesoderm [31] and involved in immunity and blood cell turnover; and gut and liver, both originating from the endoderm [32] and playing pivotal roles in digestion and metabolism. These observations emphasize the critical role of tissue-specific gene regulation in preserving the unique functionality of different tissues.

Furthermore, we pinpointed 1886 ELSs that remained active throughout development, 1334 ELSs constantly active across all tissues examined, and 702

ELSs constantly active in all biosamples surveyed (Supplementary Table 1). Such ELSs could serve as promising targets for robust transgenesis experiments. The *Tol2* transposon system [33], commonly used in zebrafish for transgenesis, randomly integrates foreign DNA into the genome, which may occur in inactive regions, potentially silencing the transgenic DNA fragment. However, our constantly active ELSs may provide advantageous sites for targeted gene knock-in experiments, ensuring robust expression.

The ELSs we've identified in this study offer a versatile and pragmatic toolkit for transgenesis in zebrafish, enabling more refined genetic manipulations. The stage- or tissue-specific ELSs provide an avenue for the controlled expression of transgenes at distinct temporal and spatial resolutions, facilitating a detailed study of gene functions and their underlying regulatory networks. Meanwhile, the constantly active ELSs present a mechanism for robust, broad-range transgene expression, making them advantageous for experiments necessitating continuous transgene activity across multiple tissues and developmental stages.

Upon examining the genomic context of these ELSs, we found that most tissue-specific ELSs were located within intronic (51%) or intergenic regions (23%) (Figure 4D and Supplementary Figure 4), consistent with a recent mouse study [34]. Notably, testis-specific ELSs exhibited the smallest proportion of intronic locations and the highest proportion within promoter regions.

We conducted a motif enrichment analysis to identify over-represented TF binding motifs in tissue-specific ELSs, inferring pivotal TFs that shape the transcriptional state of each tissue. Ranking TFs based on their -$\log_{10}$(*P*-value) of enrichment for each tissue, we noted in heart tissue (Figure 4E), for instance, that six of the top ten enriched TFs - Mef2c, Mef2d, Mef2a, GATA3, Gata6, and Gata4 - have known associations with heart function. Many top-enriched TFs in other tissues have been reported as critical for the respective tissue's function (Supplementary Figure 5). This highlights the utility of TF binding site enrichment analysis in tissue-specific ELSs for identifying key TFs driving tissue-specific gene expression.

Moreover, the genes neighboring these tissue-specific ELSs might also play significant roles in the corresponding tissue. Gene ontology analysis revealed that genes closest to tissue-specific ELSs were enriched for tissue-related functions. For example, heart-specific ELS-neighboring genes exhibited enrichment for heart processes, heart contraction, and cardiac muscle tissue development (Figure 4F). In the brain, enrichment was found for functions involved in neuron differentiation and axon development, such as neuron projection morphogenesis and cell morphogenesis. In the liver, genes for cellular response to endogenous stimulus and monocarboxylic acid metabolic process were enriched. In the testis, cilium organization and mitotic cell cycle process were enriched (Supplementary Figure 6). This underscores the notion that tissue-specific regulatory elements often reside near critical genes for relevant tissue formation and function, and bind to TFs with corresponding tissue-related regulatory roles. This further demonstrates that defining and analyzing ELS could provide valuable information about the regulatory mechanisms in each tissue.

Super-enhancers (SEs), clusters of enhancers interacting with numerous TFs, cofactors, and histone modifications, regulate gene expression and play a critical role in cell fate determination and disease progression [35, 36]. To gain insight into the key characteristics of these genomic regions, we identified SEs in the zebrafish with H3K27ac datasets and the ROSE algorithm [35] (Supplementary Figure 7). We discovered an average of approximately 470 SEs in each studied embryonic stage and adult tissue (Supplementary Figure 8A). In (Figure 4G) we illustrate a series of super-enhancers, covering a large genomic region characterized by intense ATAC-seq and H3K27ac signals, reflective of areas with high regulatory activity. The median length of the SEs we identified was 28,344 bp (Supplementary Figure 8B), in contrast to the median length of ELSs, which was 297 bp. On average, each SE contained two ELSs, and 86% of SEs contained at least one ELS (Figure 4H). GO annotation also affirmed the functional relevance of these SEs to their respective tissues. For instance, the heart-specific SEs were associated with heart morphogenesis, vasculogenesis, and cardiocyte differentiation. These SEs can be accessed through the database's genome browser.

We evaluated the *in vivo* activity of these predicted ELSs using a reporter assay, focusing on the heart. Of the 8,525 heart-activated ELSs identified, 2,494 were heart-specific. We selected 23 ELSs (Supplementary Table 2), cloned them into the reporter vector, microinjected them into zebrafish embryos, and observed the GFP signal at 3 days post-fertilization (dpf). We found that 65% (15 out of 23, including ELS22) of the ELSs drove GFP expression in the heart (Figure 5), indicating their enhancer activity during zebrafish heart development. The remaining eight ELSs did not show a GFP signal, potentially because our prediction used multi-omic data collected from various time periods. In contrast, the GFP signal was only observed at 3 dpf. This relatively high success rate (65%) of our enhancer predictions in driving GFP expression in the heart highlights the importance of integrative omics data analyses for understanding regulatory mechanisms.

Exploring multi-omics data integration

Gene regulation is a highly complex process that relies on the coordinated interaction of many regulatory factors, such as chromatin structure, TFs, regulatory elements, and their target genes. Due to the three-dimensional structure formed by complex folding and

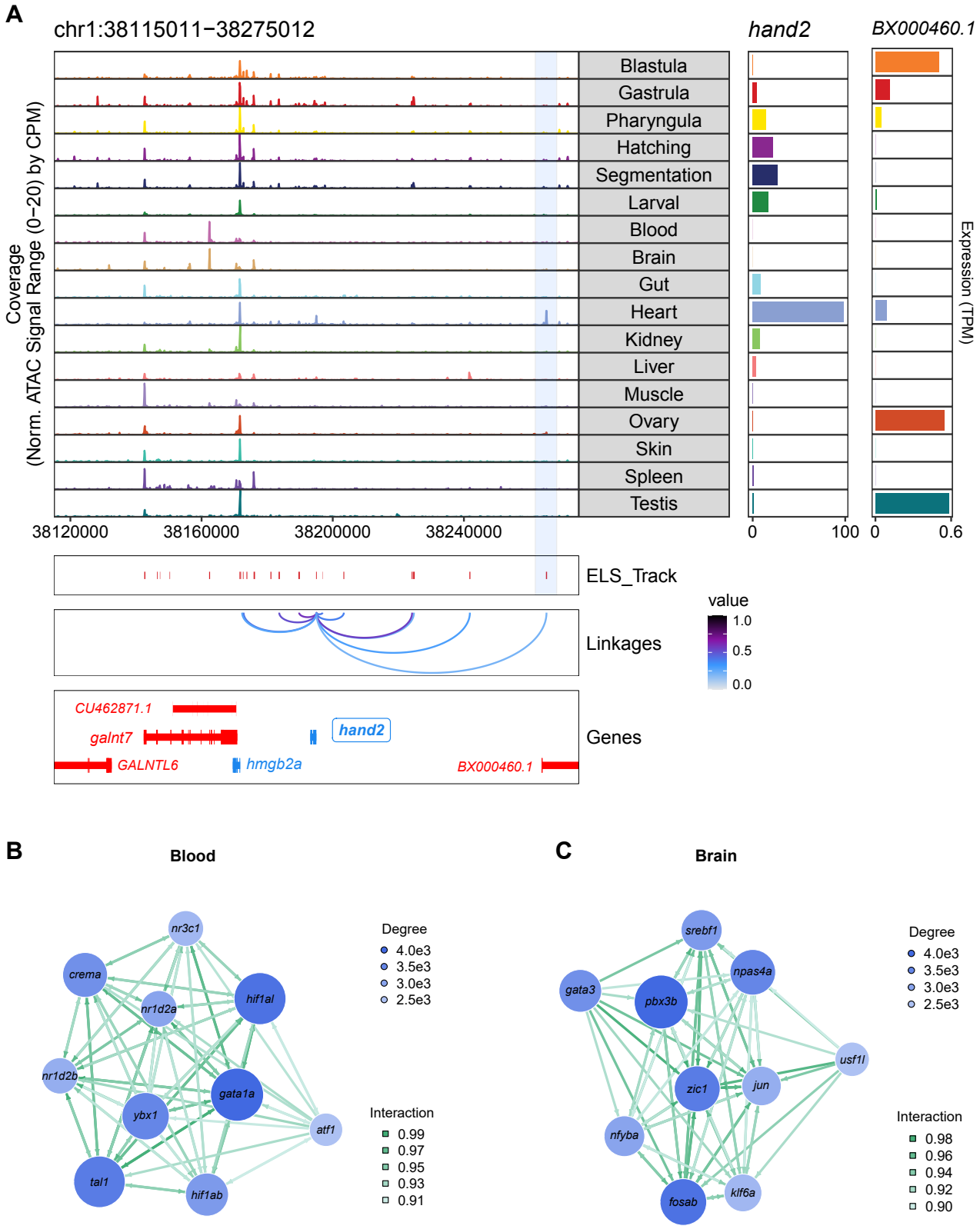

Figure 6: *Multi-omics data analysis. (A) Multi-omics plot of hand2. The central panel shows aggregated ATAC-seq genomic tracks from various samples. The right panel indicates gene expression levels in different samples computed from RNA-seq data. The bottom panel displays ELSs, predicted ELS-to-gene linkages, and genes. Notably, hand2 is predominantly expressed in the heart, with a corresponding ELS (ELS22, blue box), located 70kb downstream of hand2, specifically open in heart tissue, suggesting a linkage. (B, C) Tissue-specific enhancer-gene regulatory sub-networks for blood (B) and brain (C) tissues. Blue circles represent the top 10 TFs in each tissue, ranked by outdegree in the top 100,000 edges. The size and color intensity of the circles indicates the number of target genes for the corresponding TF. The shade of the blue arrows corresponds to the interaction score between pairs of TFs.*

assembly, interactions between regulatory elements and genes are often based on spatial rather than linear distance. Here, we estimated ELS to target gene linkages as previously reported in single-cell data [37]. As described above, we have predicted 52,350 ELSs by integrating ATAC-seq and H3K27ac ChIP-seq data. To assign these ELSs to target genes, we computed the peak matrix from ATAC-seq data and gene expression matrix from RNA-seq data in all samples. We assigned these ELSs to 26,354 target genes based on correlations between the chromosomal accessibility of ELS and the expression of the gene. Linkage correlation is a direct and interpretable measure of the linkage strength ranging from -1 to 1. All ELSs to gene linkages can be visualized as a sophisticated multi-omics plot in miniODP (Figure 6A).

Our *in vivo* ELS validation revealed one example, ELS22, capable of driving heart-specific expression (Figure 5). Located 70 kb downstream of *hand2* and specifically accessible in heart tissue, ELS22 has been linked to *hand2* based on a positive correlation between ELS22 accessibility and *hand2* expression (Figure 6A). Notably, *hand2* is recognized as crucial for heart development and regeneration [38]. Interestingly, ELS22 is located within the intron of the gene *BX000460*, which shows weak expression in the ovary and testis but not the heart. This demonstrates that the accurate linkage between ELSs and genes should be established by multi-omics evidence, not just location.

In addition to linking ELSs to genes, we predicted GRNs for various developmental stages or tissues using ANANSE [39]. We first constructed a zebrafish TF binding motif database using GimmeMotifs [40], then calculated the binding probability of all TFs based on motif scores, ATAC-seq signal, and H3K27ac signal in ELSs. Subsequently, we predicted the TF-gene binding score considering the enhancer intensity and motif score within 100 kb of the target gene's transcription start site (TSS). Finally, we ranked and scaled the binding score, TF activity, TF expression level, and target gene expression level from 0 to 1. The mean of these four scores was defined as the interaction score for the TF-gene pair, forming the GRNs.

To evaluate the quality of zebrafish developmental stage and tissue GRNs, we ranked TFs in each GRN by their outdegrees, the total number of interacting genes for each TF in each sample, and listed the top 10 TFs for all samples. Many of these are known to be critical for their respective tissues. For instance, in the blood GRN, 80% (8 out of 10) are well-known hematopoiesis factors, such as *hif1al*, *gata1a*, *hif1ab*, *tal1*, *crema*, *nr1d2a*, *nr1d2b*, *nr3c1* (Figure 6B). In the brain GRN, 60% (6 out of 10) are recognized neural regulators, such as neuronal PAS domain protein 4a (*npas4a*), brain development factor *zic1*, *pbx3b*, *gata3*, *klf6a*, *jun* (Figure 6C).

Interestingly, some vital TFs for blood development like *hif1ab* [41] and *nr3c1* [42] were not specifically expressed in blood but showed high expression in skin, spleen, kidney, brain, among other tissues (Supplementary Figure 9). However, integrated multi-omics analyses revealed that they exhibited more interactions with other genes in the blood (Figure 6B). Similarly, the TF *jun*, critical for brain development [43], showed high expression not only in the brain but also in the skin, testis, heart, and gut (Supplementary Figure 9). Yet, multi-omics analyses demonstrated that it had more connections with other genes in the brain (Figure 6C). These results highlight that TFs with higher interactions with other genes in the regulatory network play a crucial role in tissue function, reinforcing the need for integrating multi-omics rather than relying on a single omics method for predictions.

Discussion

Our study demonstrated the power and applicability of the miniENCODE analysis, employing zebrafish as a representative model. We integrated and analyzed datasets, both published and newly generated by us, from three experimental assays (RNA-seq, ATAC-seq, and ChIP-seq of H3K27ac), that covered six developmental stages and 11 adult tissues. We developed the miniODP platform, a computational tool for generating, integrating, and visualizing the vast array of multi-omics data tracks and plots. With this miniENCODE analysis of zebrafish, we were able to compile a registry of 52,350 candidate ELSs, and identify major transcription factors and their regulatory networks that were activated during these developmental stages and in the tissues. The validity of these findings, corroborated by experimental validation and existing literature, underscores the practicality and efficiency of the miniENCODE approach for understanding regulatory mechanisms in organisms. Using selected omics technologies and the reuse of published datasets, augmented by the miniODP platform, a single lab can gain valuable insights into regulatory mechanisms with relatively limited resources. This approach not only streamlines omics data interpretation but also catalyzes hypothesis generation, and enhances experimental design efficiency, which could be of considerable benefit to researchers in many fields.

The importance of CREs in gene expression, and consequently, in deciphering regulatory mechanisms, cannot be overstated. The ENCODE project exemplifies a comprehensive approach to charting genome activities, employing a broad spectrum of technologies. However, to pinpoint CREs, ENCODE ultimately relied on a more streamlined set: a strong DNase signal and at least one strong ChIP-seq signal (H3K4me3, H3K27ac, or CTCF) [2]. To further reduce the methodological intricacy and cost, our study focused solely on the DNA accessibility signal coupled with the H3K27ac histone modification signal, both hallmarks of ELSs, a critical component among all types of CREs.

The accuracy and efficiency of this approach were confirmed by our *in vivo* reporter assays. Furthermore, we embraced the evolution in the technology landscape by replacing DNase-seq with ATAC-seq [28] for capturing the chromatin accessibility and substituting ChIP-seq with the CUT&Tag method [12] for detecting histone modifications. Thus, we demonstrated that by using a pair of simplified yet powerful technologies, researchers could accurately delineate ELS in their biological samples of interest.

While the incorporation of two epigenomic profiling methods can efficiently identify ELSs, they are not sufficient to elucidate intricate regulatory mechanisms in detail. Gene expression information is required for a more comprehensive understanding, hence the inclusion of RNA-seq as a component in our essential miniENCODE assays. By integrating data from these three assays, we could link ELSs to their target genes using the correlation between gene expression and ELS activity, predict GRNs and subsequently identify key regulators within the given biological sample. Notably, many top TFs identified in the GRN analysis have been corroborated by existing literature to play critical roles in the specific tissues, further underscoring the power of integrating multi-omics datasets. This outcome reaffirms the efficacy and power of our essential miniENCODE assays in unraveling gene regulatory mechanisms.

While the essential miniENCODE assays are effective in mapping CREs and identifying regulatory mechanisms, optional miniENCODE assays could offer added depth to multi-omics analysis. Long-read RNA-seq assists in building accurate gene models, especially in newly sequenced genomes. With its superior resolution, Micro-C provides insights into long-range interactions between DNA elements. Additionally, single-cell RNA-seq, ATAC-seq, and CUT&Tag technologies allow the exploration of regulation mechanisms at the cellular level in complex tissues. Despite being more complex and costly, these optional assays can enhance the understanding of gene regulation, providing a more detailed perspective based on specific research objectives. Integrating multi-omics data stands at the forefront of genome regulation studies, providing a comprehensive view of genomic activities across diverse aspects. Such integration unveils the intricate dynamics of genomic regulation, and offers an enriched landscape of GRNs and the molecular machinery orchestrating them.

The ever-growing omics datasets generated by various research communities have highlighted an urgent need for a common computational platform to facilitate data reuse and integrated analysis. Currently, such omics platforms are predominantly limited to a few model organisms. Even for zebrafish, a widely used model organism, such an omics database is lacking. The vast accumulation of omics data can be daunting for experimental biologists searching for specific information. Our miniODP addresses this challenge, offering a user-friendly platform equipped with multiple components: the genome browser enables biologists to explore genome activities of specific loci across hundreds of data tracks, derived from different biosamples and technologies; the gene browser consolidates detailed gene-specific information from multiple external databases, facilitating the creation of plots from dozens of studies or multi-omics analyses. In essence, miniODP presents a comprehensive computational platform that unlocks the potential of published data, establishing itself as an invaluable tool for data integration and representation in omics research.

The experimental methods and database utilized in this study are generally applicable to a wide range of organisms. The complexity of the technologies utilized is manageable for individual laboratories, and the associated costs remain within an accessible range. Importantly, the miniENCODE project as outlined in this work, could serve as a catalyst for community-motivated projects. For smaller research communities, laboratories could synergize their efforts to conduct a scaled-down version of the ENCODE project, focusing on profiling the critical stages or tissues of their chosen models. By using the concept of miniENCODE outlined in this work, such a collaborative endeavor, even when conducted by a modest number of laboratories, can rapidly uncover key regulatory information about the model organism under study. This approach would empower scientific communities to expedite their research without overextending their resources, highlighting the adaptability and potential of the miniENCODE project in driving genomics research forward. In the spirit of collaboration, we extend an invitation to fellow researchers interested in exploring potential joint ventures. Our team stands ready to assist in analyzing miniENCODE datasets from various model organisms and is committed to utilizing our existing miniODP database to serve a broader scientific community.

Despite the valuable insights gained from our study, the lack of data is a clear limitation. The ENCODE project annotated the cCREs in the human and mouse genomes, representing 7.9% and 3.4% of their genomes, respectively, and accounting for 41% and 25% of the DNase-seq signal peaks [2]. The human EpiMap project annotated 2.1 million enhancers which covering 13% of the genome [44]. In contrast, our miniENCODE project of zebrafish annotated merely 1.2% of the genome as ELSs, encompassing only 17% of the ATAC-seq signal peaks. The lower proportion in zebrafish may be attributed to the limited coverage of developmental stages and tissues, lower sequencing depth, and the absence of data from more technologies for identifying other types of CREs (e.g., ChIP-seq of CTCF and H3K4me3). This underscores the need for more extensive data collection to fully unlock the regulatory landscape of the zebrafish genome.

In conclusion, our study highlights the potential and

practicality of the miniENCODE analysis in deciphering complex regulatory networks using zebrafish as a model. We demonstrated that leveraging both published and newly generated datasets, along with our newly developed miniODP computational platform, can yield comprehensive insights into gene regulation mechanisms with limited resources. Nevertheless, our understanding of the zebrafish genome's entire regulatory landscape is still emerging, and more extensive data are required. Our work serves as a starting point for future studies aiming to unravel complex regulatory mechanisms in understudied model organisms. It advocates the importance of a resource-efficient approach in genomic research, propelling advancements in diverse fields from developmental biology to disease modeling.

Methods

Zebrafish husbandry

The TU strain of zebrafish used in this study were raised and bred following standard fish-keeping procedures [45]. The fish were kept in freshwater at a temperature of $28\,^{\circ}\mathrm{C}$ and subjected to a 14-hour light / 10-hour dark cycle.

Sample collection

To isolate the ovaries of adult female fish, we first anesthetized them with Tricaine methanesulfonate (MS222) and then placed them in a 2% agarose groove. Next, we used ophthalmic scissors and fine forceps to open the abdomen and transfer the ovaries to a tube containing 2 mL of DPBS (Gibco, cat. C14190500BT). We minced the ovaries into <1 mm pieces using microdissection scissors to obtain small pieces of tissue. Once the tissue had sunk to the bottom of the tube, we replaced the medium with 2 mL of digestive enzyme mixture I (3 mg/mL collagenase I (Worthington, cat. LS004196), 3 mg/mL collagenase II (Worthington, cat. LS004174), and 1.6 mg/mL hyaluronidase (Worthington, cat. LS002594) in Leibovitz L-15 Medium (Gibco, cat. 21083027)). We incubated the mixture at $37\,^{\circ}\mathrm{C}$ with 15 revolutions per minute (RPM) on an orbital rotator. Every 10 minutes, we monitored the suspension until no or a minimal number of cell clumps were observed. We then centrifuged the cells for 5 minutes at 300 *g*, and resuspended them in 2 mL of digestive enzyme mixture II (0.25% TRYPSIN-EDTA (Gibco, cat. 15400-054), 0.8 U/mL DNase I (Invitrogen, cat. M0303S) in L-15), and incubated them for 10 minutes at $37\,^{\circ}\mathrm{C}$ on an orbital rotator. We terminated the trypsin reaction by adding $200\,\mu\mathrm{L}$ of FBS (Sigma-Aldrich, cat. F8318). The cell pellet was then resuspended, washed twice with 1 mL of DPBS containing 0.04% BSA, and centrifuged at 500 *g* for 5 minutes. Finally, we filtered the cell pellet using a $100\,\mu\mathrm{m}$ cell strainer.

To ensure that all embryos were at the same developmental stage, we followed the criteria outlined by Kimmel [46] and checked the diagnostic characteristics of each embryo individually. We eliminated chorions from the embryos by incubating them in 1 mg/mL Pronase for 10 minutes, followed by washing in DPBS. To remove the yolks, we incubated the embryos in 1 mL of deyolking buffer (55 mM NaCl, 1.8 mM KCl and 1.25 mM $NaHCO_3$) and pipetted 20-30 times. The embryo tissues were homogenized in 2 mL of FACSmax cell dissociation solution (Genlantis, cat. T200100) and then incubated at room temperature for 5 minutes. After passing the cells through a $40\,\mu\mathrm{m}$ cell strainer, we centrifuged them at 310 *g* for 5 minutes. We measured cell density and vitality using the Countstar Rigel S2.

Library preparation

ATAC-seq library The samples were dissected freshly and processed for ATAC-seq immediately. Briefly, the tissues were first resuspended in 1.5 mL of lysis buffer (10 mM Tris-HCl (pH 7.4), 10 mM NaCl, 3 mM $MgCl_2$, 0.10% Tween-20 (Bio-rad, cat. 1662404EDU), 0.10% NP-40 (Abcam, cat. ab142227), 0.01% Digitonin, 1% BSA in Nuclease-free Water). Following filtration through a $40\,\mu\mathrm{m}$ cell strainer, nuclei were collected by centrifugation at 500 g for 5 minutes. ATAC-seq libraries were prepared as previously described [47] with minor modifications. After removing the supernatant, the nuclei were resuspended in $50\,\mu\mathrm{L}$ of transposition mix ($10\,\mu\mathrm{L}$ 5x TTBL, $5\,\mu\mathrm{L}$ TTEmix V50 (Vazyme, cat. TD501), $0.5\,\mu\mathrm{L}$ 1% digitonin, $0.5\,\mu\mathrm{L}$ 10% Tween-20, $34\,\mu\mathrm{L}$ ddH_2O). Following the PCR procedure, 0.6x AMPure beads (Beckman, cat. A63880) were used to eliminate large fragments, and 1.2x AMPure beads were used to recover the library. In this study, the Illumina Novaseq platform was used to sequence all ATAC-seq libraries, each of which included two duplicates.

CUT&Tag library CUT&Tag was performed following established protocols [12] with minor modifications. In brief, the Hyperactive Universal CUT&Tag Assay Kit for Illumina (Vazyme, cat. TD903) was used to generate libraries. Cells were harvested and washed twice with 1.5 mL of Wash Buffer using gentle pipetting. Next, $10\,\mu\mathrm{L}$ of activated concanavalin A-coated magnetic beads were added and incubated at room temperature for 10 minutes. Cells bound to magnetic beads were resuspended in $50\,\mu\mathrm{L}$ of chilled Antibody Buffer containing $1\,\mu\mathrm{L}$ of H3K27ac antibody (Abcam, cat. ab4729) or Rabbit IgG (Abcam, cat. ab171870) primary antibody, followed by overnight incubation at $4\,^{\circ}\mathrm{C}$. The primary antibody was removed by placing the tube on the magnet stand until the solution became clear and the supernatant was removed. Cells were incubated at room temperature for 60 minutes with Guinea Pig anti-Rabbit IgG (Heavy & Light Chain)

antibody (Antibodies-Online, cat. ABIN101961). To remove binding antibodies, cells were washed with the magnet stand twice for 5 minutes in 0.2 mL of Dig-Wash buffer. With gentle vortexing, 50 µL of 0.04 µM pA/G-Tn5 adapter complex in Dig-300 Buffer was added to the cells, which were then incubated at room temperature for 1 hour. Cells were washed in 0.2 mL of Dig-300 Buffer. Thereafter, cells were resuspended in 50 µL of Tagmentation buffer (10 µL of 5x TTBL in 40 µL of Dig-300 Buffer) and incubated at 37 °C for 60 minutes. To stop tagmentation and extract DNA from 50 µL of sample, 5 µL of Proteinase K, 100 µL of Buffer L/B, and 20 µL of DNA Extract Beads were added. The beads were washed twice in 200 µL of Buffer WA and 200 µL of Buffer WB. After the beads were allowed to dry, 22 µL of ddH_2O was added. To amplify libraries, 15 µL of DNA was mixed with 5 µL of a universal i5, 5 µL of a uniquely barcoded i7 primer, and 25 µL of 2x CAM. Finally, 100 µL of VAHTS DNA Clean Beads (Vazyme, cat. N411) were added to purify the library.

Enhancer reporter assays

To evaluate the biological relevance of putative tissue-specific ELSs, we employed a *Tol2* transposon-mediated zebrafish transgenesis approach, as previously described [48]. In brief, selected enhancers were PCR amplified from zebrafish genomic DNA and subcloned into the pSTG vector upstream of the pig minimum EF1a promoter-EGFP cassette (see Supplementary Table 2 for primer details). These enhancer reporter plasmids were then co-injected with transposase mRNA into TU zebrafish embryos at the one-cell stage. Fluorescence microscopy (Leica) was used for whole embryo visualization and image capture at 3 dpf, focusing on cardiac-specific regions. At least 100 embryos were examined for each ELS.

Data analysis

Data processing We collected genome and high-throughput omics data of zebrafish and medaka to construct a multi-omics database. To ensure the quality of the raw data, FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used, and low-quality sequences were filtered out using Trim Galore (http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) prior to the downstream analysis. A standard bioinformatics workflow was applied to all RNA-seq datasets, including alignment with HISAT2 (version 2.2.1) [49] and quantitative analysis of all genes using Stringtie (version 2.1.2) [50]. Transcripts per kilobase million reads (TPM) were used as the metric for quantifying gene expression.

For the ChIP-seq and ATAC-seq omics data, we aligned reads to the zebrafish (GRCz11) or medaka reference genome downloaded from Ensembl using Bowtie2 (version 2.4.2) [51]. Duplicate fragments were filtered using SAMtools [52] and Picard (http://broadinstitute.github.io/picard/), and peaks in the ATAC-seq and ChIP-seq data were called using MACS2 (version 2.2.7.1) [53]. We indexed the generated SAM/BAM format alignments using SAMtools. To assess the quality of epigenomic samples, we computed three quality control measures: insert-size distribution, TSS enrichment score, and Pearson correlation. The TSS enrichment score represents the ratio between the average sequencing depth/reads count centered on TSSs and that was flanking the TSSs.

For WGBS, Bisulfite-Seq, and other methylation-related data sets, we aligned and converted them into coverage using Bismark (version 0.22.3) [54] and deepTools (version 3.4.3) [55]. The resulting Bigwig files can be accessed through JBrowse for all analyzed data.

For scRNA-seq data, we utilized either the gene expression count table downloaded directly or the raw data processed by Cell Ranger (version 4.0.0) as input for Seurat (version 3.2.2) [56]. Subsequently, we filtered the data based on nFeature and nCount presence and mitochondrial content, as reported in the original papers. In the case of snATAC-seq data, we constructed a gene activity score matrix. We performed cluster analysis based on the top 2000 variable genes and computed principal components (PCs) for UMAP or t-SNE plots using ElbowPlot and JackStraightPlot. We then employed the FindAllMarkers function to identify differentially expressed genes in each cluster.

Interactive visualization tools JBrowse [20] was integrated into miniODP to enable the display of annotated genomic features, genome sequences, and other genomic information for all available reference genome assemblies. JBrowse is particularly effective when dealing with large piecemeal features such as variations. The gene viewer tools were implemented using the R Shiny package (https://shiny.rstudio.com/) and ShinyProxy (https://www.shinyproxy.io/), with an SQLite database as the backend. ShinyProxy is deployed to enhance concurrency. The BLAST tool in miniODP was implemented using SequenceServer [24], which offers a user-friendly interface with interactive visual outputs.

Enhancer-like signatures (ELSs) prediction of zebrafish We utilized multi-omics data and followed the methodology described by the ENCODE project [2] to predict potential enhancers. Specifically, we employed representative DNA accessibility region (rDAR) identification. To address the variation in sequencing depths across ATAC-seq datasets, we utilized MACS2 to identify ATAC-seq peaks. We subsequently merged adjacent regions with bedtools merge to create ATAC peaks, discarding any sections shorter than 50 bp or longer than 500 bp. We kept the most significant peak and removed overlapping peaks from further analysis. This procedure was iterated on the remaining peaks until no additional peaks remained.

For each rDAR, we computed the Z-scores of the log10-transformed ATAC and H3K27ac signals in each biosample. To account for the histone modifications at the flanking nucleosomes, we averaged the H3K27ac signals throughout a larger region, including the rDAR plus 500 bp on either side. We defined ELS as rDAR with a high ATAC max-Z and another high max-Z for H3K27ac ChIP-seq marks. We employed a Z-score threshold of 1.64, corresponding to the 95th percentile of a one-sided Gaussian distribution. A high Z-score or maximum Z-value was defined as >1.64 for each base pair within the range, while a low score was defined otherwise. For group classification, we also considered the genomic distance between the ELS's center and the nearest TSS (0-2000 bp for TSS-proximal and >2000 bp for TSS-distal). ELSs were classified as either TSS-proximal enhancer-like signatures (pELS) or TSS-distal enhancer-like signatures (dELS).

ELS to gene linkages To estimate ELS-to-gene linkages, we utilized a peak matrix and a gene expression matrix, as previously reported in single-cell data [37]. First, we identified ELS, as described above. Second, we computed ATAC-seq reads in the ELS matrix and created the sample x ELS matrix. Similarly, we computed RNA-seq read in the gene matrix and generated the sample x gene matrix. Third, we identified all feasible ELS-to-gene pairs within a 250 kb window surrounding the TSS. Then, we computed the Pearson correlation of the log2-normalized values between the samples x ELS matrix and the sample x gene matrix for all samples. Additionally, subsets of ELS-to-gene links can be selected and visualized in miniODP.

GO and motif analysis Gene Ontology enrichment analysis was performed to identify enriched biological processes and molecular functions using clusterProfiler (version 4.0.5). Additionally, we predicted enriched motifs for each sample using HOMER [57].

Identification of super-enhancers The H3K27ac ChIP-seq datasets were aligned to the GRCz11 reference genomes. The ROSE (Rank Ordering of Super-Enhancers) algorithm (https://bitbucket.org/young_computation/rose) was used to distinguish super-enhancers from enhancers [35].

Gene regulatory network (GRN) analysis ANANSE [39] can build a GRN using gene expression, ATAC-seq, and ChIP-seq data. The Network Influence Score can rank transcription factors (TFs) according to their likelihood of regulating crucial genes for cell reprogramming. To build a motif database for zebrafish, we aggregated all vertebrate motifs from the CIS-BP database with GimmeMotifs [40] and converted them using the motif2factors function. In ELS, both the ATAC and H3K27ac signals were measured, and these signals were used in combination with motif scores to forecast TF binding using the ananse binding function. The resulting predicted TF binding sites and the expression levels of both TFs and their target genes were used to construct a sample-specific GRN. Four metrics were used to construct TF-gene interaction scores and network edge weights: (1) the predicted TF binding probability in enhancers, (2) the distance between enhancers and the target gene, (3) the expected TF activity, (4) the expression level of both TF and the target gene. Finally, the results were visualized using the igraph R package (version 1.3.1).

Code and demonstration files: For the convenience of readers who wish to apply these analyses to their own datasets, we have developed a Docker image as the running environment. This image, along with the source code, documentation, and demonstration files, is accessible via our GitHub repository at https://github.com/QTuLab/miniENCODE. By utilizing these resources, readers can conduct the analyses with the provided demo files and rapidly establish a functional miniODP platform, following the step-by-step instructions included.

Data and code availability

ATAC-seq and CUT&Tag data are deposited in the NCBI Gene Expression Omnibus (GEO) database under accession GSE207155 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE207155, reviewer token obmtikycxzunfgz). The miniODP database is available at: https://tulab.genetics.ac.cn/miniodp/. Code, along with documentation, a configured running environment, and example files for both data analysis and data visualization, can be found at: https://github.com/QTuLab/miniENCODE.

References

1. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* **306,** 636–640 (2004).
2. ENCODE Project Consortium *et al.* Expanded Encyclopaedias of DNA Elements in the Human and Mouse Genomes. *Nature* **583,** 699–710 (2020).
3. modENCODE Consortium *et al.* Identification of Functional Elements and Regulatory Circuits by Drosophila modENCODE. *Science* **330,** 1787–1797 (2010).
4. Gerstein, M. B. *et al.* Integrative Analysis of the Caenorhabditis Elegans Genome by the modENCODE Project. *Science* **330,** 1775–1787 (2010).
5. Lieschke, G. J. & Currie, P. D. Animal Models of Human Disease: Zebrafish Swim into View. *Nat Rev Genet* **8,** 353–367 (2007).
6. Yang, H. *et al.* A Map of Cis-Regulatory Elements and 3D Genome Structures in Zebrafish. *Nature* **588,** 337–343 (2020).
7. Howe, D. G. *et al.* The Zebrafish Information Network: Major Gene Page and Home Page Updates. *Nucleic Acids Res* **49,** D1058–D1064 (2021).

8. Tan, H., Onichtchouk, D. & Winata, C. DANIO-CODE: Toward an Encyclopedia of DNA Elements in Zebrafish. *Zebrafish* **13,** 54–60 (2016).
9. Baranasic, D. *et al.* Multiomic Atlas with Functional Stratification and Developmental Dynamics of Zebrafish Cis-Regulatory Elements. *Nat Genet* **54,** 1037–1050 (2022).
10. Petryszak, R. *et al.* Expression Atlas Update–an Integrated Database of Gene and Protein Expression in Humans, Animals and Plants. *Nucleic Acids Res* **44,** D746–752 (2016).
11. Papatheodorou, I. *et al.* Expression Atlas Update: From Tissues to Single Cells. *Nucleic Acids Res* **48,** D77–D83 (2020).
12. Kaya-Okur, H. S. *et al.* CUT&Tag for Efficient Epigenomic Profiling of Small Samples and Single Cells. *Nat Commun* **10,** 1930 (2019).
13. Sharon, D., Tilgner, H., Grubert, F. & Snyder, M. A Single-Molecule Long-Read Survey of the Human Transcriptome. *Nat Biotechnol* **31,** 1009–1014 (2013).
14. Lieberman-Aiden, E. *et al.* Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* **326,** 289–293 (2009).
15. Hsieh, T.-H. S. *et al.* Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell* **162,** 108–119 (2015).
16. Tang, F. *et al.* mRNA-Seq Whole-Transcriptome Analysis of a Single Cell. *Nat Methods* **6,** 377–382 (2009).
17. Buenrostro, J. D. *et al.* Single-Cell Chromatin Accessibility Reveals Principles of Regulatory Variation. *Nature* **523,** 486–490 (2015).
18. Wu, S. J. *et al.* Single-Cell CUT&Tag Analysis of Chromatin Modifications in Differentiation and Tumor Progression. *Nat Biotechnol* **39,** 819–824 (2021).
19. Bartosovic, M., Kabbe, M. & Castelo-Branco, G. Single-Cell CUT&Tag Profiles Histone Modifications and Transcription Factors in Complex Tissues. *Nat Biotechnol* **39,** 825–835 (2021).
20. Buels, R. *et al.* JBrowse: A Dynamic Web Platform for Genome Visualization and Analysis. *Genome Biol* **17,** 66 (2016).
21. Cunningham, F. *et al.* Ensembl 2022. *Nucleic Acids Res* **50,** D988–D995 (2022).
22. Hu, H. *et al.* AnimalTFDB 3.0: A Comprehensive Resource for Annotation and Prediction of Animal Transcription Factors. *Nucleic Acids Res* **47,** D33–D38 (2019).
23. Paysan-Lafosse, T. *et al.* InterPro in 2022. *Nucleic Acids Res* **51,** D418–D427 (2023).
24. Priyam, A. *et al.* Sequenceserver: A Modern Graphical User Interface for Custom BLAST Databases. *Mol Biol Evol* **36,** 2922–2924 (2019).
25. Barrett, T. *et al.* NCBI GEO: Archive for Functional Genomics Data Sets–Update. *Nucleic Acids Res* **41,** D991–995 (2013).
26. Leinonen, R. *et al.* The European Nucleotide Archive. *Nucleic Acids Res* **39,** D28–31 (2011).
27. Gasperini, M., Tome, J. M. & Shendure, J. Towards a Comprehensive Catalogue of Validated and Target-Linked Human Enhancers. *Nat Rev Genet* **21,** 292–310 (2020).
28. 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).
29. Creyghton, M. P. *et al.* Histone H3K27ac Separates Active from Poised Enhancers and Predicts Developmental State. *Proc Natl Acad Sci U S A* **107,** 21931–21936 (2010).
30. Heinz, S., Romanoski, C. E., Benner, C. & Glass, C. K. The Selection and Function of Cell Type-Specific Enhancers. *Nat Rev Mol Cell Biol* **16,** 144–154 (2015).
31. Prummel, K. D., Nieuwenhuize, S. & Mosimann, C. The Lateral Plate Mesoderm. *Development* **147,** dev175059 (2020).
32. Zorn, A. M. & Wells, J. M. Vertebrate Endoderm Development and Organ Formation. *Annu Rev Cell Dev Biol* **25,** 221–251 (2009).
33. Suster, M. L., Kikuta, H., Urasaki, A., Asakawa, K. & Kawakami, K. Transgenesis in Zebrafish with the Tol2 Transposon System. *Methods Mol Biol* **561,** 41–63 (2009).
34. Sarropoulos, I. *et al.* Developmental and Evolutionary Dynamics of Cis-Regulatory Elements in Mouse Cerebellar Cells. *Science* **373,** eabg4696 (2021).
35. Whyte, W. A. *et al.* Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. *Cell* **153,** 307–319 (2013).
36. Hnisz, D. *et al.* Super-Enhancers in the Control of Cell Identity and Disease. *Cell* **155,** 934–947 (2013).
37. Granja, J. M. *et al.* ArchR Is a Scalable Software Package for Integrative Single-Cell Chromatin Accessibility Analysis. *Nat Genet* **53,** 403–411 (2021).
38. Xin, M., Olson, E. N. & Bassel-Duby, R. Mending Broken Hearts: Cardiac Development as a Basis for Adult Heart Regeneration and Repair. *Nat Rev Mol Cell Biol* **14,** 529–541 (2013).
39. Xu, Q. *et al.* ANANSE: An Enhancer Network-Based Computational Approach for Predicting Key Transcription Factors in Cell Fate Determination. *Nucleic Acids Res* **49,** 7966–7985 (2021).
40. van Heeringen, S. J. & Veenstra, G. J. C. GimmeMotifs: A de Novo Motif Prediction Pipeline for ChIP-sequencing Experiments. *Bioinformatics* **27,** 270–271 (2011).
41. Gerri, C., Marass, M., Rossi, A. & Stainier, D. Y. R. Hif-1α and Hif-2α Regulate Hemogenic Endothelium and Hematopoietic Stem Cell Formation in Zebrafish. *Blood* **131,** 963–973 (2018).
42. Kwan, W. *et al.* The Central Nervous System Regulates Embryonic HSPC Production via Stress-Responsive Glucocorticoid Receptor Signaling. *Cell Stem Cell* **19,** 370–382 (2016).
43. Han, R. *et al.* Trim69 Regulates Zebrafish Brain Development by Ap-1 Pathway. *Sci Rep* **6,** 24034 (2016).
44. Boix, C. A., James, B. T., Park, Y. P., Meuleman, W. & Kellis, M. Regulatory Genomic Circuitry of Human Disease Loci by Integrative Epigenomics. *Nature* **590,** 300–307 (2021).
45. Westerfield, M. *The Zebrafish Book. A Guide for the Laboratory Use of Zebrafish (Danio Rerio)* 4th ed (Univ. of Oregon Press, Eugene, 2000).
46. Kimmel, C. B., Ballard, W. W., Kimmel, S. R., Ullmann, B. & Schilling, T. F. Stages of Embryonic Development of the Zebrafish. *Dev Dyn* **203,** 253–310 (1995).

47. Wu, J. *et al.* Chromatin Analysis in Human Early Development Reveals Epigenetic Transition during ZGA. *Nature* **557,** 256–260 (2018).
48. Xie, W. *et al.* Epigenomic Analysis of Multilineage Differentiation of Human Embryonic Stem Cells. *Cell* **153,** 1134–1148 (2013).
49. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: A Fast Spliced Aligner with Low Memory Requirements. *Nat Methods* **12,** 357–360 (2015).
50. Pertea, M. *et al.* StringTie Enables Improved Reconstruction of a Transcriptome from RNA-seq Reads. *Nat Biotechnol* **33,** 290–295 (2015).
51. Langmead, B. & Salzberg, S. L. Fast Gapped-Read Alignment with Bowtie 2. *Nat Methods* **9,** 357–359 (2012).
52. Li, H. *et al.* The Sequence Alignment/Map Format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
53. Zhang, Y. *et al.* Model-Based Analysis of ChIP-Seq (MACS). *Genome Biol* **9,** R137 (2008).
54. Krueger, F. & Andrews, S. R. Bismark: A Flexible Aligner and Methylation Caller for Bisulfite-Seq Applications. *Bioinformatics* **27,** 1571–1572 (2011).
55. Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A. & Manke, T. deepTools: A Flexible Platform for Exploring Deep-Sequencing Data. *Nucleic Acids Res* **42,** W187–191 (2014).
56. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177,** 1888–1902.e21 (2019).
57. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime Cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol Cell* **38,** 576–589 (2010).

Acknowledgements

We thank Dr. Jing-Wei Xiong (Peking University) for providing the TU strain of zebrafish; Dr. Tong Chen (State Key Laboratory for Quality Ensurance and Sustainable Use of Dao-di Herbs) and Mr. Minghong Zhang (Institute of Genetics and Developmental Biology) for help with website development. This work was supported by the Chinese Academy of Sciences (CAS) Project for Young Scientists in Basic Research (YSRB073), CAS Key Research Program (ZDBS-ZRKJZ-TLC009), and National Natural Science Foundation of China (32200439).

Author contributions

Q.T. and Y.L. designed the project. H.Y. performed most experiments and computational analysis. Z.S. and H.S. assisted with the experiments. Y.L. and P.J. assisted with computational analysis. Q.T., H.Y., and Y.L. wrote the manuscript.

Additional information

Supplementary Information: accompanies this paper at Supplemental Figures and Supplementary Table 1 and 2.

Competing interests: The authors declare no competing interests.