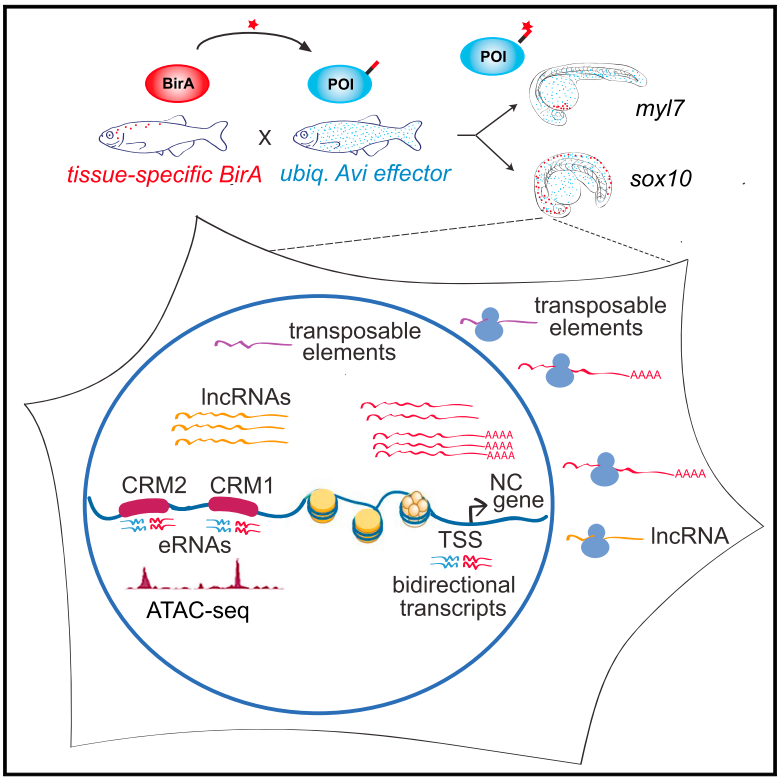
Cell Reports

Resource

Biotagging of Specific Cell Populations in Zebrafish Reveals Gene Regulatory Logic Encoded in the Nuclear Transcriptome

Graphical Abstract



Highlights

- Biotagging enables cell- and compartment-specific in vivo biotinylation in zebrafish
- Technique yields comprehensive nuclear transcriptional analysis of cardiomyocytes
- Biotagging finds bidirectionally transcribed neural crest *cis*-regulatory modules
- System reveals tissue-specific regulation of noncoding RNA species

Authors

Le A. Trinh, Vanessa Chong-Morrison, Daria Gavriouchkina, Tatiana Hochgreb-Hägele, Upeka Senanayake, Scott E. Fraser, Tatjana Sauka-Spengler

Correspondence

tatjana.sauka-spengler@imm.ox.ac.uk

In Brief

A genetically encoded in vivo biotinylation system in zebrafish developed by Trinh et al. reveals cell-type- and subcellular-compartment-specific coding and non-coding RNAs in developing cardiomyocytes and neural crest cells. Characterization of non-coding RNAs in neural crest reveals bidirectionally transcribed *cis*-regulatory elements that define a specific gene regulatory signature.

Accession Numbers

GSE89670

Trinh et al., 2017, Cell Reports *19*, 425–440
April 11, 2017
http://dx.doi.org/10.1016/j.celrep.2017.03.045

CellPress

Cell Reports
Resource

Biotagging of Specific Cell Populations in Zebrafish Reveals Gene Regulatory Logic Encoded in the Nuclear Transcriptome

Le A. Trinh,[1,4] Vanessa Chong-Morrison,[2,4] Daria Gavriouchkina,[2] Tatiana Hochgreb-Hägele,[2,3] Upeka Senanayake,[2] Scott E. Fraser,[1] and Tatjana Sauka-Spengler[2,5,*]
[1]Molecular and Computational Biology, University of Southern California, Los Angeles, CA 90089, USA
[2]Radcliffe Department of Medicine, Weatherall Institute of Molecular Medicine, University of Oxford, Oxford OX3 9DS, UK
[3]Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91125, USA
[4]Co-first author
[5]Lead Contact
*Correspondence: tatjana.sauka-spengler@imm.ox.ac.uk
http://dx.doi.org/10.1016/j.celrep.2017.03.045

SUMMARY

Interrogation of gene regulatory circuits in complex organisms requires precise tools for the selection of individual cell types and robust methods for biochemical profiling of target proteins. We have developed a versatile, tissue-specific binary in vivo biotinylation system in zebrafish termed biotagging that uses genetically encoded components to biotinylate target proteins, enabling in-depth genome-wide analyses of their molecular interactions. Using tissue-specific drivers and cell-compartment-specific effector lines, we demonstrate the specificity of the biotagging toolkit at the biochemical, cellular, and transcriptional levels. We use biotagging to characterize the in vivo transcriptional landscape of migratory neural crest and myocardial cells in different cellular compartments (ribosomes and nucleus). These analyses reveal a comprehensive network of coding and non-coding RNAs and *cis*-regulatory modules, demonstrating that tissue-specific identity is embedded in the nuclear transcriptomes. By eliminating background inherent to complex embryonic environments, biotagging allows analyses of molecular interactions at high resolution.

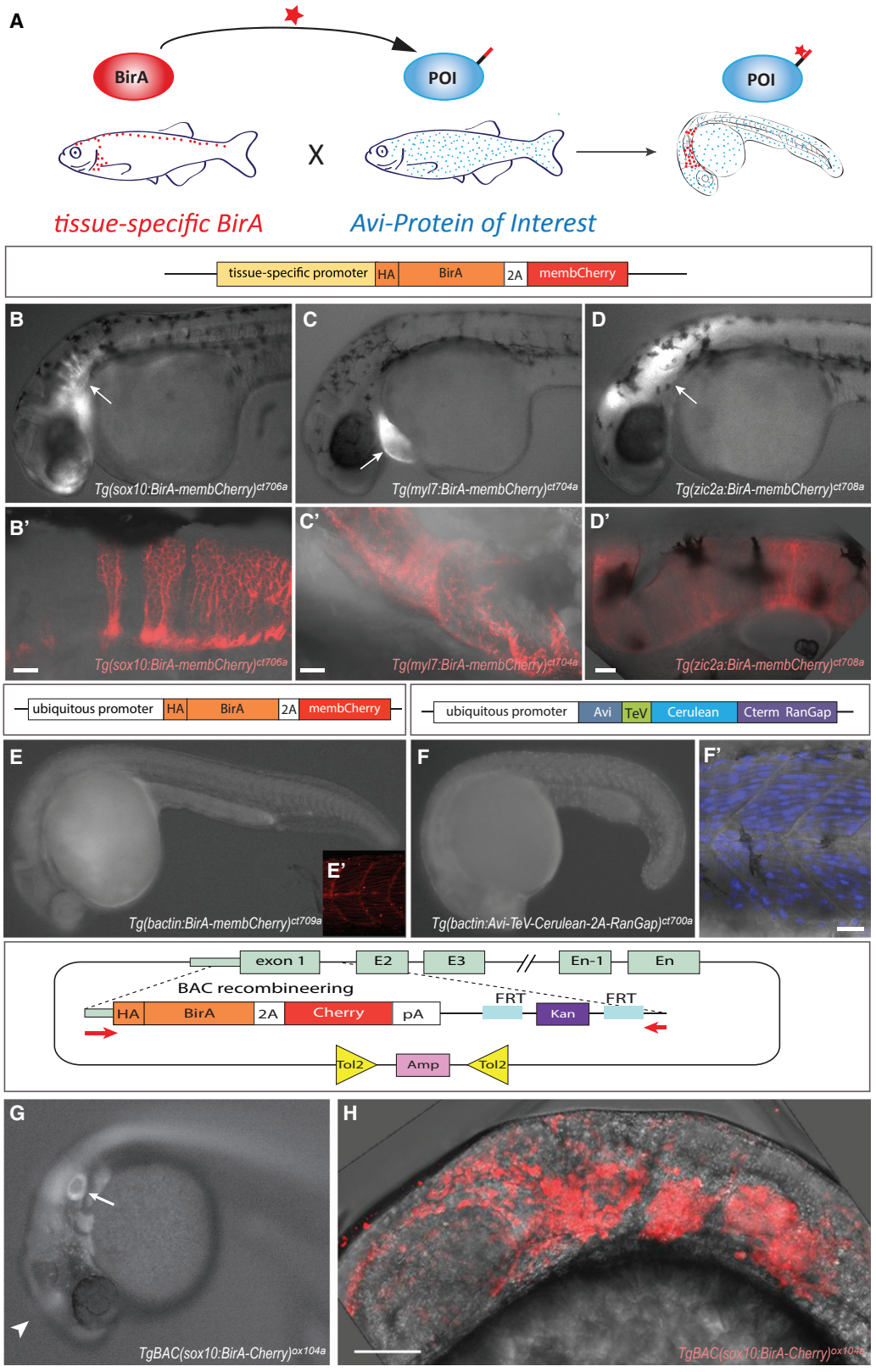
INTRODUCTION

Multicellular organisms are a complex mixture of cell types, each within a unique microenvironment and exposed to different cell interactions that result in the execution of distinct transcriptional programs. This complicates analyses of gene regulatory networks, since intermingled cell types are often present in small numbers. Moreover, subcellular RNA localization provides a supplementary level of control. Such issues highlight the need for the efficient isolation of defined subcellular compartments of individual cell populations from their in vivo context in the organism and optimized genome-wide regulatory profiling protocols applicable to small samples.

Current cell isolation approaches in vertebrates have a number of drawbacks for such analyses. Laser microdissection and fluorescence-activated cell sorting (FACS) can isolate subpopulations but require specialized equipment and involve lengthy processing times, during which cell state and gene expression can change. Expanding cell numbers in culture is risky, as the cellular microenvironments are not easily recapitulated in vitro. Isolating subcellular compartments requires lengthy fractionation procedures that can further alter the sample or degrade signals. In vivo biotinylation circumvents these limitations, and a number of strategies have been employed to isolate subcellular compartments for transcriptional, chromatin or proteomic profiling in plants and animals (Amin et al., 2014; Deal and Henikoff, 2010; Ooi et al., 2010; Steiner et al., 2012). These approaches involve co-expression of biotin ligase (BirA) and a biotin acceptor peptide (Avi tag) fused to a protein of interest (Cronan, 1990; de Boer et al., 2003). Because the biotin-avidin interaction is one of the strongest non-covalent interactions in nature ($K_d \sim 10^{-15}$), this approach permits streptavidin-based affinity purification of protein targets and their interacting entities (e.g., nucleic acids, proteins, and entire nuclei) with high stringency.

Isolation of nuclei tagged in specific cell types (INTACT) involves biotinylation of an Avi-tagged fusion protein that binds to the nuclear envelope for affinity purification of nuclei (Deal and Henikoff, 2010), allowing active transcriptome profiling and studies of chromatin features. In vivo biotinylation of Avi-tagged Rpl10 protein in zebrafish embryos can purify ribosomes via the translating ribosome affinity purification (TRAP) method (Heiman et al., 2008) for translational profiling (Housley et al., 2014). A full understanding of the RNA landscape and its regulation would require profiles of both subcellular compartments.

We sought to exploit the power of in vivo biotinylation in zebrafish and generate a genetic binary system for biotin labeling of subcellular compartments in different tissue-specific contexts. To simplify the nomenclature, we collectively termed the



(legend on next page)

labeling, purification, and analysis approach "biotagging." The biotagging toolkit consists of two types of transgenic lines: (1) BirA drivers that express biotin ligase in a tissue-specific manner and (2) a set of Avi-effectors expressing zebrafish-compatible versions of Avi-tagged proteins used for INTACT and TRAP. Combining different biotagging driver and effector lines, we optimized procedures for specific biotinylation and stringent isolation of defined subcellular compartments for cell-type-specific epigenomic, transcriptional, and proteomic profiling in zebrafish. By comparing genome-wide regulatory profiles obtained from nuclei and ribosomes in migrating neural crest (NC), developing myocardium, and whole embryos, we identified developmentally regulated and tissue- and subcellular compartment-specific RNAs that include protein coding and long non-coding RNAs (lncRNA) and transposable elements. Furthermore, we uncovered divergent (bidirectional) transcription of active enhancers and promoters.

We establish the utility of the biotagging approach by performing chromatin accessibility assays and quantitative tissue-specific analysis of enhancer transcription in the nuclei of migrating NC, permitting us to identify and rank NC-specific enhancers. Our results highlight the molecular basis of tissue-specific gene regulatory networks encrypted in the nuclear transcriptome, revealed by nascent transcription across both coding and non-coding regions. Our genetic toolkit and analysis pipelines permit investigation of gene regulatory circuits and molecular phenotyping at the systems level in specific cell types in vivo.

RESULTS AND DISCUSSION

Building the Biotagging Toolkit

Drawing on the power of zebrafish genetics, the biotagging toolkit was created as a modular system, encoding the components needed for specific biotinylation in separate transgenic lines, so it can be tailored to any cell population of interest and genetic background of choice. Using transposon-mediated transgenesis and bacterial artificial chromosome/clone (BAC) recombineering, we generated sets of biotinylation "driver" lines (seven tissue-specific and four ubiquitous lines) that reliably express BirA (Figures 1 and S1; Table S1) and five "effector" lines expressing Avi-tagged target proteins (Figures 1F, 1F′, and 2; Table S1). When Avi-effector fish are crossed with BirA driver lines, biotinylation of the target protein occurs only in embryos that carry both transgenes and only in cells that co-express both components (Figure 1A).

The biotagging toolkit supports the isolation of nuclei via INTACT (Deal and Henikoff, 2010) or ribosomes via TRAP (Heiman et al., 2008; Tryon et al., 2013) through Avi-effector lines that add an Avi tag and a fluorescent label to each subcellular compartment. The effectors (nucAvi and riboAvi) use beta-actin2 (*βactin*) or ubiquitin (*ubiq*) promoters to drive ubiquitous expression of a zebrafish-compatible Avi-Cerulean-RanGap or Avi-Cerulean-Rpl10 fusion protein, tagging the outer nuclear envelope or ribosomes, respectively (see Supplemental Experimental Procedures; Figure S1; Table S1). Imaging of the nucAvi or riboAvi lines confirmed localization of effector proteins on nuclei (Figures 1F′, 2C, 2D, 3A, 3B, and S1) or in cytoplasm (Figures 2E and 2F).

Optimizing and Testing the Biotinylation Parameters of Biotagging in Zebrafish

To assess the specificity and selectivity of BirA for Avi-tagged proteins in zebrafish, we performed immunoblotting of protein extracts of embryos from crosses of NC-specific BirA driver lines, ncBirA and ncBirA(BAC) (Figures 1B and 1G), with either the nucAvi(*βactin*) or riboAvi(*ubiq*) effector lines. Expression of BirA in the driver lines did not lead to biotinylation of endogenous proteins over the background level observed in wild-type embryos (Figures 3C and 3D, lanes 1 and 4), even when BirA was overexpressed (Figure S2A, lane 2). Similarly, the Avi tag remained non-biotinylated by zebrafish endogenous biotin ligases (Figure 3C, lane 2). Efficient biotinylation was achieved without supplementation with biotin in embryos carrying both Avi-effector and BirA-driver alleles (Figures 3C and 3D). To define minimal expression requirements for the biotagging approach, we studied samples that have low expression of either of the components. We found that a low level of BirA was sufficient for effective biotinylation, but a low effector level resulted in decreased biotinylation of the Avi tag (Figure S2).

Isolation of Total RNA from Nuclei and Ribosomes in Selected Cell Types

Co-expression of BirA and nucAvi or riboAvi enables efficient isolation of biotinylated nuclei or ribosomes using streptavidin magnetic beads (Figures 3E–3E″; see Experimental Procedures). In a direct comparison of different total RNA isolation protocols (biotagged nuclei, biotagged ribosomes, and FACS), biotagging the nuclei of NC cells resulted in a ~7-fold higher yield per embryo over the FACS approach; biotagging ribosomes was ~5-fold better (Figure 3F). Bioanalyzer profiles revealed that nuclear total RNA is distinct from ribosomal and whole-cell total RNA profiles (Figures 3G–3I), with a broader range of sizes and a significantly smaller fraction of 18S and 28S rRNAs (Figure 3G; ~5% of total nuclear RNA versus ~50% of whole-cell

Figure 1. Genetically Encoded Biotagging Toolkit in Zebrafish

(A) Schematic of the binary transgenic system for cell-type-specific in vivo biotinylation. BirA drivers are in red and Avi effector lines in blue. POI, protein of interest.

(B–E) Widefield fluorescent image of biotagging drivers expressing BirA under the *sox10* (B), *myl7*(C), and *zic* (D), or ubiquitous (E, *βactin*) promoters; schematics of the transgenic constructs are shown above the images. (B′–E′) Corresponding confocal images of BirA-equivalent membCherry expression in the pharyngeal arches and hindbrain (white arrow, B), myocardium (white arrow, C), and hindbrain (white arrow, D).

(F) Widefield image of Avi-RanGap(nucAvi) effector with schematic of transgenic construct above image. (F′) Confocal image of Avi-RanGap expression in the somite.

(G and H) Widefield fluorescent (G) and projection of confocal (H) microscope images of ncBirA(BAC) driver with schematic of recombineering BAC cassette and *tol2* containing BAC. Arrow points to otic vesicle, and arrowheads point to midbrain expression.

Scale bars represent 50 μm (B′–D′) and 20 μm (F′ and H).

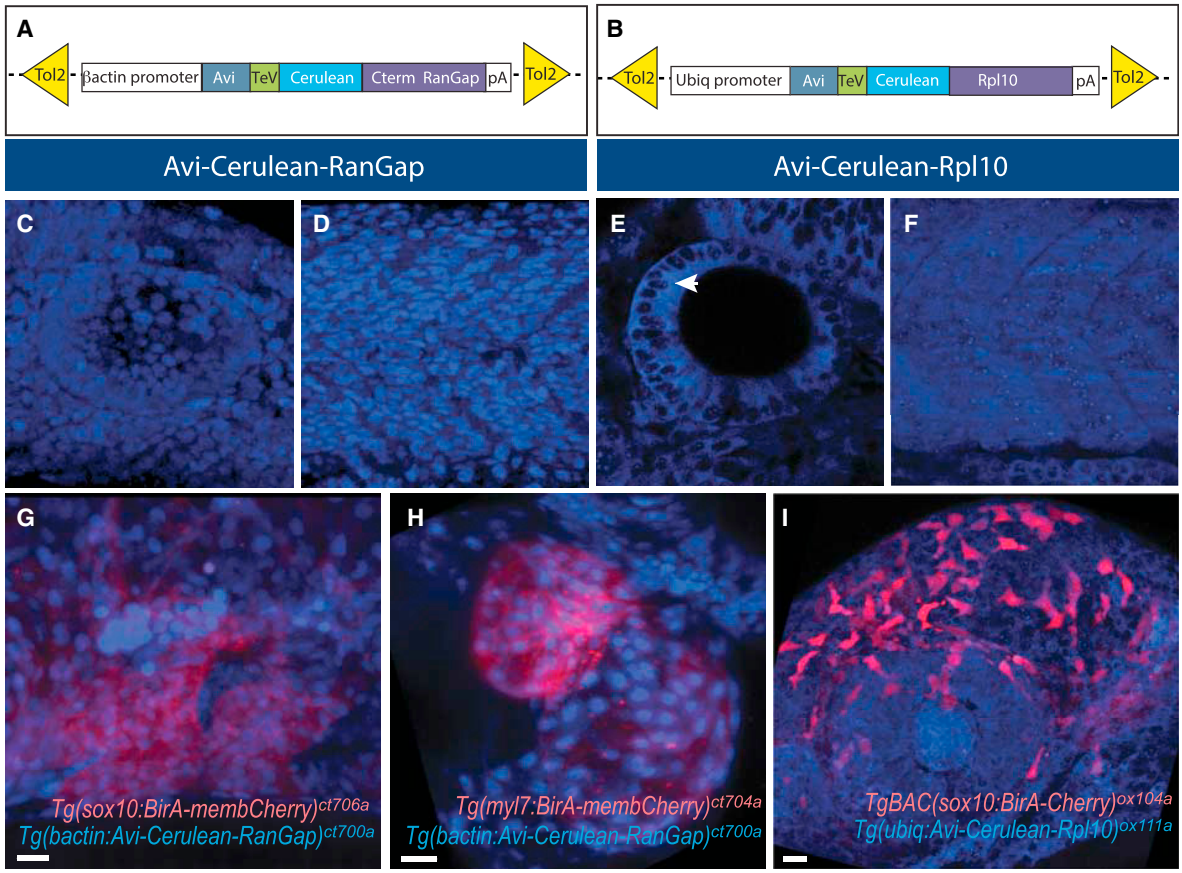


Figure 2. Biotagging Avi-Tagged Effectors

(A and B) Schematic of Avi-tagged constructs for generating nuclear effector (nucAvi) (A) and ribosome effector (riboAvi) (B).

(C–F) Confocal 3D projection of nucAvi (C and D) and riboAvi (E and F) expression in the developing inner ear (C and E) and somite (D and F) at 32 hpf. Arrow points to nucleoli.

(G–I) Confocal 3D projection of BirA driver (G and I in NC; H in myocardium, red) and Avi effector (G,H, nucAvi, and I, riboAvi, blue). Scale bars, 20μm.

total RNA) (Barthelson et al., 2007). The striking resemblance between total RNA profiles from bound (specific) and unbound (flow-through) nuclei (Figures 3G and 3H) indicates the comprehensive cellular lysis and stringency of our optimized nuclear isolation procedures (see Supplemental Experimental Procedures). Isolated NC nuclei represented ~2% of the nuclei from the whole embryos (based on fluorescence unit [FU] units level or overall RNA concentration; see Supplemental Experimental Procedures), which closely corresponds to percentage of NC cells in the embryo. The distinct RNA contents and high yields validate the use of biotagging to isolate desired subcellular compartments.

Genome-wide Analysis Validates Tissue-Specificity of Biotagging

Profiling nuclear RNA pools provides direct characterization of the active transcriptome, particularly relevant when studying gene regulatory circuitry (Mitchell et al., 2012; Zaghlool et al., 2013). To cross-validate our approach, we compared the presence of tissue-specific signatures in 26–30 hours post-fertilization (hpf) myocardial nuclei to whole-embryo nuclei (stage-matched controls) isolated from crosses of myoBirA or ubBirA(*βactin*) drivers with the nucAvi(*βactin*) effector (referred to as *myl7* and *bactin* nuclear datasets). Because many nuclear RNA species are not polyadenylated, we used ribo-depletion, rather than poly(A)-based RNA selection, and prepared strand-specific sequencing libraries (see Experimental Procedures).

Differential expression analysis comparing *myl7* and *bactin* nuclear samples identified 6,750 differentially expressed genes ($p < 0.05$), with 3,715 genes significantly enriched and 3,035 depleted in the *myl7* nuclear samples (Figure 4A). Gene set enrichment analysis (GSEA) revealed the presence of several signaling pathways implicated in cardiac development and function, such as Wnt, cadherin, and Rho GTPase-mediated pathways (Figure 4B). The largest node from the GSEA consisted of 76 Wnt pathway genes with the largest edge consisting of 24 cadherin pathway genes (Figure 4B), which is in line with previous evidence of their involvement in early heart development (Brade et al., 2006; Gessert and Kühl, 2010). Statistical over-representation analysis of the *myl7* dataset reveals enriched gene ontologies (GOs) of processes related to muscle contraction and muscle organ and mesoderm development. Furthermore,

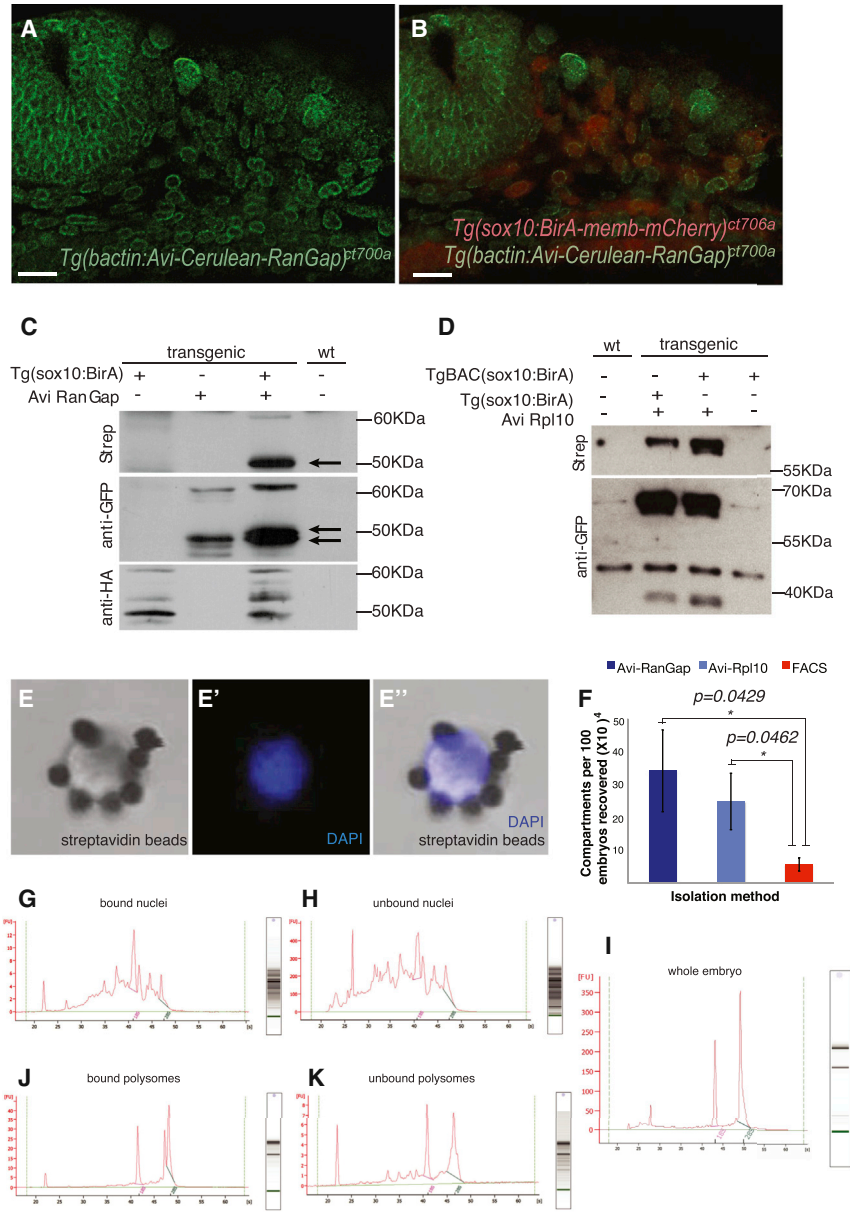


Figure 3. Specific In Vivo Biotinylation of Avi-Tag Proteins and Purification of Subcellular Compartments

(A and B) Antibody staining for Avi-RanGap (green) and HA-BirA (red), with anti-GFP and anti-hemagglutinin (anti-HA) antibodies, respectively. In fixed samples, Avi-RanGap localize more discretely to the nuclear envelope. Anti-HA staining shows BirA (red) expressed in both nuclei and cytoplasm of cells. Scale bars, 20 μm.

(C and D) Streptavidin, anti-GFP, and anti-HA western blot of nuclear (C) and ribosome (D) extracts from BirA drivers (ncBirA or ncBirA(BAC)) and Avi-tagged (nucAvi, C) or (riboAvi, D) effector embryos. (C) Arrow points to biotinylated Avi-RanGap (C, lane 3), shifted to larger size after biotinylation when detected with anti-GFP (compare lanes 2 and 3).

(E) Bright-field image of harvested nucleus from BirA;nucAvi embryos after incubation with streptavidin Dynabeads and isolated by magnetic capture. (E′) DAPI stained of nucleus in (E). (E″) Merge of images in (E) and (E′).

(F) Quantification of total RNA yield from biotagged nuclei or ribosomes or FACS isolation protocols using ncBirA and respective Avi-tagged effectors. RNA from cellular compartments calculated per 100 embryos. Error bars represent SDs from two sequenced replicates. Significance calculated using Student's t-test (one-tailed, two-sample equal variance).

(G–K) Representative Bioanalyzer profile of total RNA extracted from Streptavidin-bound biotagged nuclei (G), ribosomes (J), flow-through (unbound) (H and K), and whole embryo (I).

enriched protein class GOs included essential regulators of cardiovascular function such as actin family cytoskeletal proteins, actin-binding proteins, and G protein modulators (Figure 4C). Surveying the ZFIN expression database (Bradford et al., 2011), we found that 357 of 419 annotated myocardial genes were expressed in the *myl7* nuclear datasets at 2 FPKMs or higher. A statistically significant number of those (133/419, $p < 0.01$) were overrepresented in *myl7* versus *bactin* nuclei (Figures S3A and S3B; Table S2). Differential expression analysis of the *myl7* and migratory NC datasets (17–18 hpf) confirms their divergence, with known myocardial genes showing enrichment in *myl7* nuclei (Figure 4D). Independent assays of RNA enrichment by qPCR of *myl7* nuclei show a 6- to 13-fold enrichment of myocardial genes (Figure S3C). Biotagging nuclear profiling is highly reproducible, recovering the cardiomyocyte transcriptional signature with low variance between replicates (Figure S3).

Strand-Specific Profiling of NC Nuclear RNA Reveals Pervasive Transcription at Open Loci and Cell-Type-Specific Divergent Transcription

Differential expression of ribo-depleted total RNA from NC and whole-embryo nuclei (16–18 somite stage [ss]; 17–18 hpf) did not recover a clear NC signature according to gene models annotated in Ensembl (mostly protein-coding genes). However, pathways implicated in the formation of NC derivatives are revealed by differential and GO analyses of nuclear poly(A)-selected transcriptomes at a later stage (24 hpf) (Figure S4). Given that we observed prominent pervasive transcription across the genome in our early NC nuclear datasets, we reasoned that the less distinct differential expression might reflect stem-like features of the NC cells at this stage, as stem cells are characterized by indiscriminate nuclear transcription (Guenther et al., 2007). To further investigate this hypothesis and deduce the regulatory architecture that might underlie

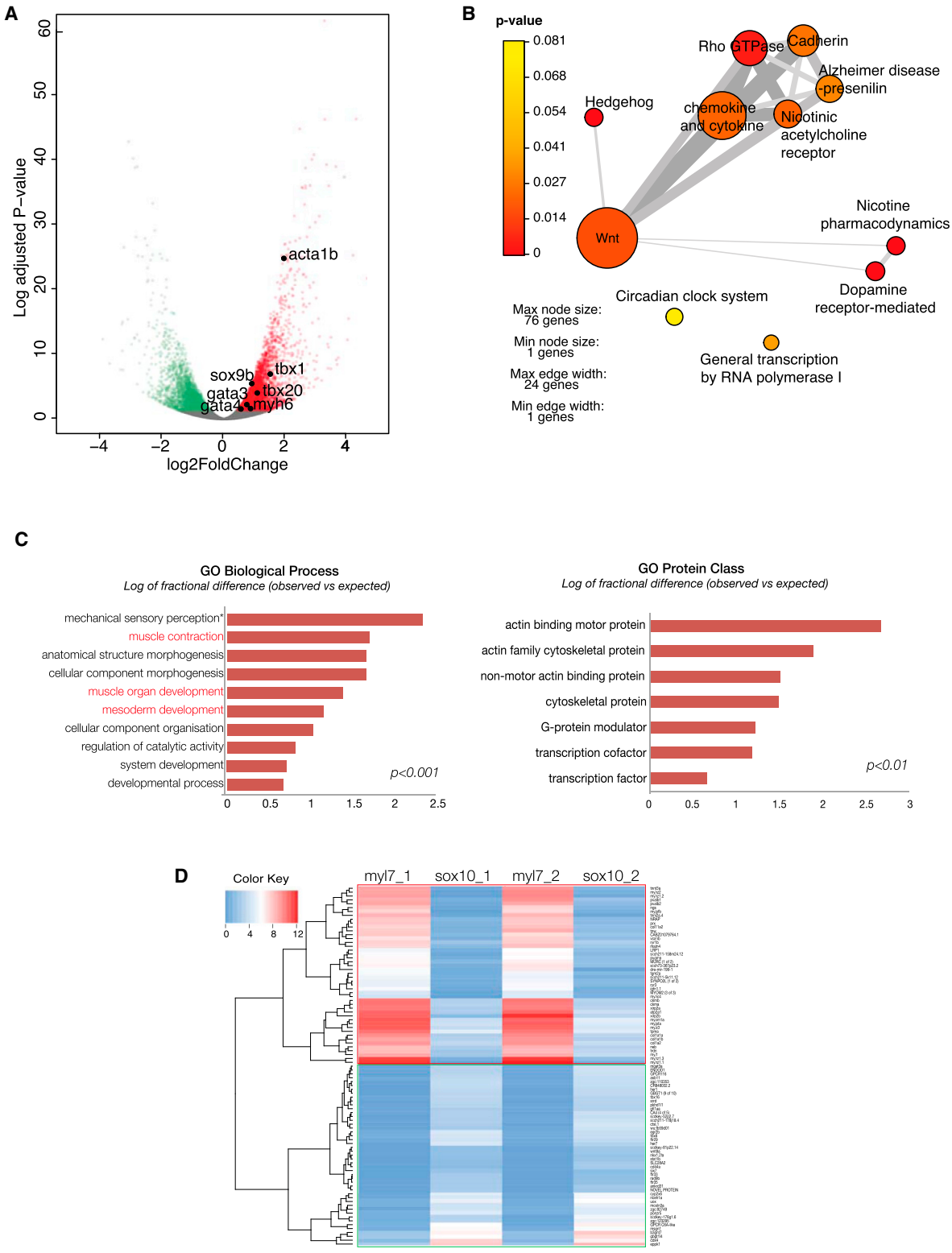


Figure 4. Enrichment of Cell-Type Signature by Biotagging

(A) Volcano plot of differential expression between *myl7* and *bactin* nuclear transcriptomes ($p < 0.05$; red, enriched; green, decreased in *myl7* samples). Black dots represent known myocardial genes.

(B) GSEA of genes enriched in *myl7* nuclear dataset. Size of node corresponds to number of genes in each gene set. p values are presented by color saturation.

(legend continued on next page)

pervasive transcription in early NC, we identified regions of accessible chromatin by assay for transposase-accessible chromatin using sequencing (ATAC-seq) (Buenrostro et al., 2013) performed on migratory NC cells isolated from ncBirA(BAC) embryos. In addition, we have used TRAP biotagging to analyze the actively translated fractions of migrating NC cells and stage-matched controls (crossing the riboAvi effector line with ncBirA(-BAC) (*sox10* ribosome) and ubBirA(*βactin*) drivers, respectively). Isolated ribosomal RNA pools were enriched using ribo-depletion and used for construction of strand-specific cDNA libraries.

The presence of short bidirectional transcripts resulting from divergent transcription initiated within the same genomic region but in opposite directions is a known hallmark of active promoters (Core et al., 2008; Guenther et al., 2007; Seila et al., 2008). We used our strand-specific datasets to compare divergent transcription at the active promoters in NC and whole-embryo nuclei. Open promoters (ATAC_TSS set) were defined as ATAC-seq-positive regions at the 5′ end of Ensembl-annotated zebrafish genes. To account for gene misannotation, we extended this window by 100 bp from the transcription start sites (TSSs). Quantification of our transcriptional datasets split by strands showed that open promoters were indeed pervasively transcribed (Figures 5A and 5B). In NC nuclear datasets, a majority of the 16,660 TSS ATAC peaks were transcribed (15,305 on the "+" strand and 15,323 on the "−" strand; ∼92%). The majority (86%; 14,295) of these exhibited bidirectional transcription (Figure 5B). In contrast, only ∼62% of the TSS ATAC-peaks were transcribed in the *bactin* nuclear datasets (10,414 on the + strand and 10,204 on the − strand) and only ∼32% (5,383) were transcribed bidirectionally (Figure 5B). This greater divergent transcription at TSS in NC nuclei suggests that the undifferentiated state and broad potential of migratory NC cells may be sustained by extensively open and transcribed chromatin, as proposed for stem cells (Guenther et al., 2007).

k-means clustering using linear normalization of the stranded transcription in 16–18 ss samples (NC nuclear, NC ribosomal and *bactin* nuclear) revealed ten distinct gene clusters with varying levels of short bidirectional transcripts at open promoters (Figure S5). Cluster organization reflected the coding strand direction and structural organization of a gene within the analyzed region of ±1.5 kb from TSS. We identified five clusters that assembled open promoter elements (TSS ATAC-seq peaks) and were bidirectionally transcribed in NC nuclei (clusters 1–5; Figures 5C and 5D). Scatterplot quantification of normalized counts showed that ∼55% of these loci (1884/3391 in Cl1-3, 1657/2986 in Cl4-5) were specific to NC nuclear samples and only ∼5% (93/1,600 and 68/1,397) to *bactin* nuclear datasets. Similarly, comparison of individual enriched clusters (Cl.1-5, Figures 5E and 5F) highlighted clear differences in their Pearson correlation coefficients (Ye et al., 2011).

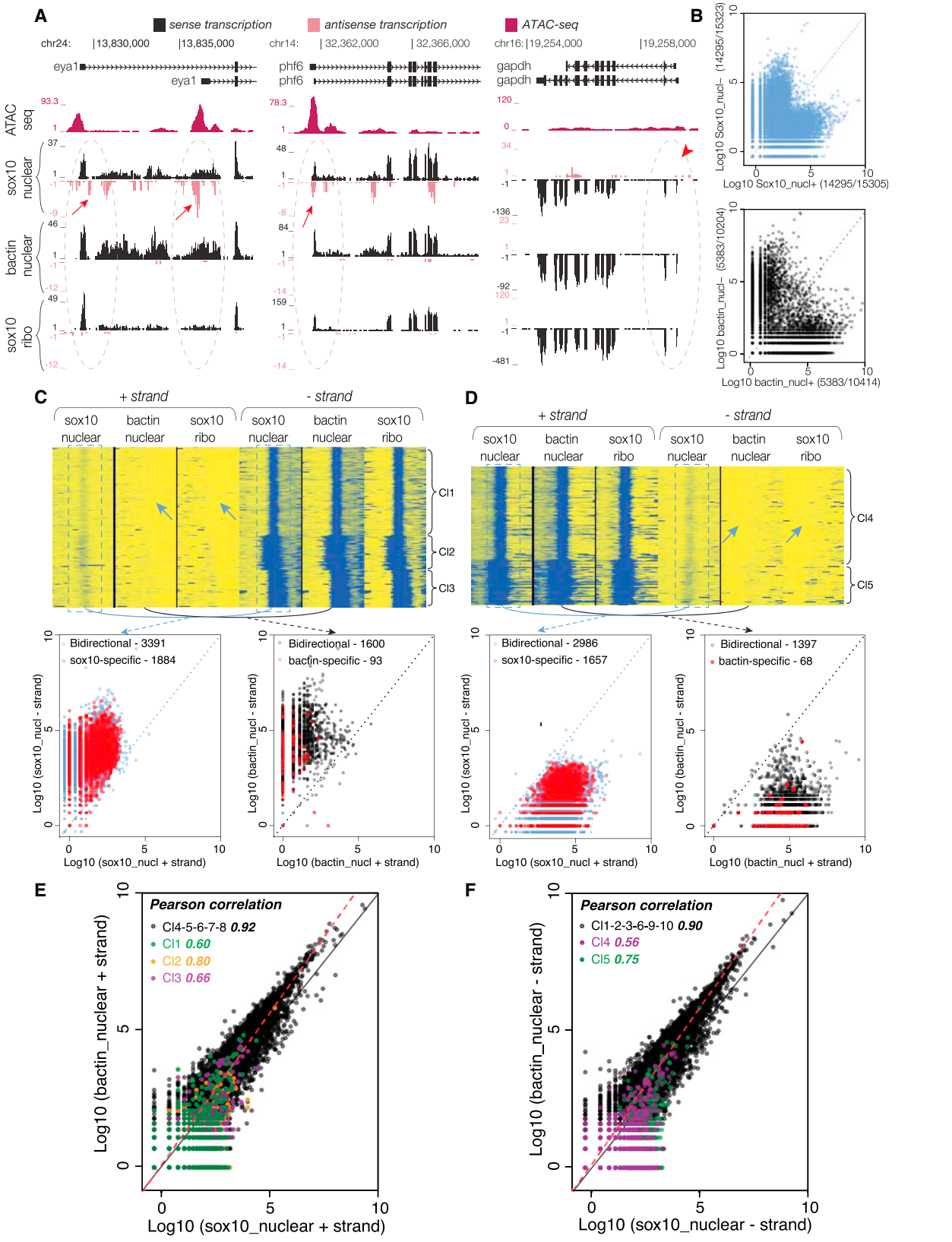
To compare the genes exhibiting bidirectional transcription and those that do not, we used statistical overrepresentation tests and GO term functional classification. The top enriched GO terms associating ($p < 0.01$) to loci with bidirectionally transcribed TSSs included developmental processes such as eye and sensory organ morphogenesis, neurogenesis, and cellular differentiation. This is in sharp contrast to the GO terms significantly enriched ($p < 0.01$) for loci not exhibiting bidirectional transcription, which reflect multiple metabolic processes (Figures S5B and S5C). When ranked according to either protein class or biological function, the most striking difference found between these gene clusters was a sharp increase in transcription factors, including all known bona fide NC and otic placode regulators among bidirectionally transcribed loci (Figure S5D). These findings are in line with previous suggestions that antisense transcription is associated with promoters of transcriptional regulators (Lepoivre et al., 2013), arising as a consequence of RNA polymerase II (Pol II) stalling (Core et al., 2008; Nepal et al., 2013). The presence of poised RNA Pol II at promoters driving important developmental regulators has been proposed to be critical for the coordination of transcriptional events during development, allowing dynamic and rapid gene activation (Boettiger and Levine, 2009; Gaertner et al., 2012; Zeitlinger et al., 2007).

Antisense transcripts at divergent promoters undergo nuclear exosome complex recruitment and degradation once mRNA transcripts are spliced and stabilized (Andersson et al., 2015; Preker et al., 2008). Thus, there is a higher chance of detecting antisense transcripts at newly activated genes than at the TSS of active loci, where Pol II stalling is thought to be absent (Hendrix et al., 2008; Zabidi et al., 2015). Interestingly, for some loci, this analysis revealed pervasive upstream antisense transcription even in the ribosomal samples, albeit at lower frequencies (clusters 2, 3, and 5; Figures 5C and 5D). We reasoned that these events most likely correspond to long non-coding RNAs (lncRNAs) that are preferentially transcribed in the vicinity of active promoters in antisense orientation (Sigova et al., 2013).

Nuclear Transcriptome Analysis Uncovers NC Cis-regulatory Elements

Similar to active promoters, associated *cis*-regulatory elements are pervasively bidirectionally transcribed, resulting in nuclear-enriched enhancer RNAs (eRNAs) (Andersson et al., 2014; Core et al., 2014; De Santa et al., 2010; Kim et al., 2010; Kowalczyk et al., 2012). These short eRNAs are sensitive to degradation by the nuclear exosome complex, much like the upstream antisense transcripts from divergent promoters of protein-coding genes (Andersson et al., 2015). Therefore, although promoters and enhancers share many unifying features (core elements, divergent transcription, and transcription factor [TF] binding), the fundamental distinction between them is the greater RNA stability of post-initiation sense RNA transcripts (Andersson et al., 2015; Core et al., 2014). Recent studies suggest that enhancer transcription correlates with outputs from the

(C) GO terms for biological processes and protein class enriched in *myl7* nuclear dataset compared to *bactin* nuclear dataset with $p < 0.001$ and $p < 0.01$, respectively. Red text indicates terms related to cardiac function.
(D) Heatmap of top 50 differentially enriched genes in either *myl7* (red framed) or *sox10* nuclear samples (green framed). Log2-fold enrichment is presented in blue-red color key.



(legend on next page)

downstream coding genes and may represent the earliest event in the gene activation cascade (Arner et al., 2015).

We used our nuclear transcriptome datasets obtained from NC and whole-embryo nuclei at 17–18 hpf to identify the ensemble of putative active enhancers coordinating the NC regulatory program and the associated NC transcriptional signature. Due to their rapid degradation, eRNAs are usually difficult to detect in relatively small samples obtained from specific cell types in vivo. Notably, our nuclear datasets are significantly enriched in eRNAs, rendering them ideally suited to this type of analysis (Figure 6A). We used NC-specific ATAC-seq to delineate a set of putative distal regulatory elements for further analysis (ATAC_enhancer set), which we defined as extragenic ATAC peaks that did not overlap with Ensembl-annotated promoter regions or exons. To determine whether NC nuclear transcriptional profiles exhibit tissue-specific patterns of enhancer transcription and identify putative *cis*-regulatory modules (CRMs), we have applied the *k*-means clustering algorithm to strand-specific datasets obtained from NC and whole-embryo nuclei using the seqMINER platform (Ye et al., 2011). Linear enrichment clustering of RNA-seq outputs was computed genome-wide over ATAC_enhancer peaks (±1.5 kb from the center) (Figure 6B). We have identified two distinct cohesive clusters of CRMs (one on each strand) with clear patterns of short eRNA bidirectional transcription in NC nuclei, but not in whole-embryo nuclear or ribosomal samples (clusters 1 and 2; 17,071 CRMs; Figure 6B). The merged profile for clusters 1 and 2 indicated a similar enrichment in divergent transcription of ATAC_enhancer regions in NC versus whole-embryo nuclear samples (Figure 6C). A third cluster (cluster 3; 2,561 CRMs) with similar "architecture" (divergent transcription in NC nuclei only; Figure 6B), included elements transcribed across longer regions surrounding the ATAC-peaks and most likely contained long intergenic non-coding RNAs (lincRNAs), transcribed transposons, and enhancers. To quantify the enrichment at ATAC_enhancer regions between NC and whole-embryo nuclear samples, we plotted the values for divergent transcription and calculated Pearson correlation coefficients for different *k*-means clusters. We show that values for "NC-specific" clusters 1–3 (R_{Cl1} = 0.23, R_{Cl2} = 0.39, and R_{Cl3} = 0.02; Figure 6D) are significantly offset from the coefficient for all clusters (R_{all} = 0.75; Figure 6D). Other identified clusters contained non-transcribed elements or "ubiquitous" elements, transcribed in both whole-embryo and NC nuclei or even detected in the ribosomal compartment (Figure S6A). Interestingly, while the median value of ATAC-seq read density on transcribed (clusters 1–3) and non-transcribed regions (cluster 4) is similar, there is a greater variation in the ATAC-seq signal for the non-transcribed elements (Figure S6B).

To study the tissue-specific activity of CRMs in clusters 1 and 2, we defined the level of NC-specific divergent transcription as the ratio (fold change [FC]) in transcriptional output (total fragments per kilobase per million mapped reads [FPKM] over ATAC_enhancer peaks) between NC (*sox10*) and whole-embryo (*bactin*) nuclear samples. Ranking the FC values for all valid CRMs (11,655 with FPKM > 1 for *NC* and *bactin*) (Figure 6E) revealed three brackets of CRM activity (low, FC < 1; intermediate, 1 < FC < 5; high, FC > 5), corresponding to different levels of tissue-specific eRNA enrichment. When annotated, we found that CRMs associated with known NC genes (NC expression at 14–19 ss according to the ZFIN in situ database) are significantly enriched in the intermediate and high FC brackets (1 < FC; $p < 0.001$), unlike CRMs associated with ubiquitously expressed or otic genes, which at these stages were not statistically significant (Figure 6E).

We used the Genomic Regions Enrichment of Annotations Tool (GREAT) (McLean et al., 2010) to test if the collection of CRMs identified as differentially transcribed in *sox10*-positive nuclei harbored a NC regulatory signature. GREAT allows assignment of functional significance to a set of non-coding genomic regions by analyzing the annotations of nearby genes and integrating statistically significant distal regulatory elements. GREAT analysis of clusters 1 and 2 (5,087 elements with 1 < FC < 5; yellow box in Figure 6E) revealed an enrichment of functional GO terms associated with biological processes related to NC and otic placode formation (Figure 6F). This reflects the expression of ncBirA(BAC) at 16–18 ss in migrating and differentiating NC cells, as well as the otic placode (Figures 1G and 1H). This highly specific enrichment of NC-associated GO terms obtained using a whole genome as background was statistically significant by both binomial and hypergeometric tests (Benjamini $p < 0.01$). Highlighted terms included NC development/migration as well as biological processes covering the entire complement of NC derivatives (e.g., glia, pigment cells, sympathetic neurons, pectoral fin mesenchyme, and adrenal gland NC contributions; Figure 6F). Therefore, the ensemble of CRMs obtained from analysis of *sox10* nuclei identifies a set of active enhancers implicated in migrating and differentiating NC in vivo.

Tight tissue-specific expression of key developmental regulators is thought to result from the combinatorial activity of multiple *cis*-regulatory elements. When annotated, expressed genes associated with NC CRMs from clusters 1 and 2 were ranked

Figure 5. Strand-Specific Nuclear RNA Profiles Reveal Divergent Transcription

(A) Genome browser screenshot illustrating antisense transcription (red arrows) at active promoters of newly actively transcribed genes (ATAC peaks, jazzberry jam), but not at the housekeeping locus *gapdh* (red arrowhead).

(B) Scatterplot of raw counts mapped to open promoter at TSS from NC (top) and *bactin* (bottom) nuclear samples split by strand (+/−). Number of TSSs with bidirectional transcripts (14,295, *nc*; 5,383, *bactin*) out of total elements in parentheses.

(C and D) Heatmap depicting *k*-means clustering of strand-specific transcription at TSSs of actively transcribed genes (3,871 elements on the "−" strand in C and 3,370 elements on the "+" strand in D) using linear normalization and corresponding raw counts scatterplots. Bidirectionally transcribed TSSs are boxed (light blue) and indicated by either blue dots in NC or black dots in *bactin* nuclear dataset. Cell-type-specific elements are indicated in red.

(E and F) Scatterplot quantification of enrichment in bidirectional transcription at TSSs between NC and *bactin* nuclei for +/− strand normalized counts. Pearson correlation coefficient (*r*) for enriched *k*-clusters 1, 2, and 3 on the + strand (E) and *k*-clusters 4 and 5 on the − strand (F) presented in different colors. Control clusters are shown in black.

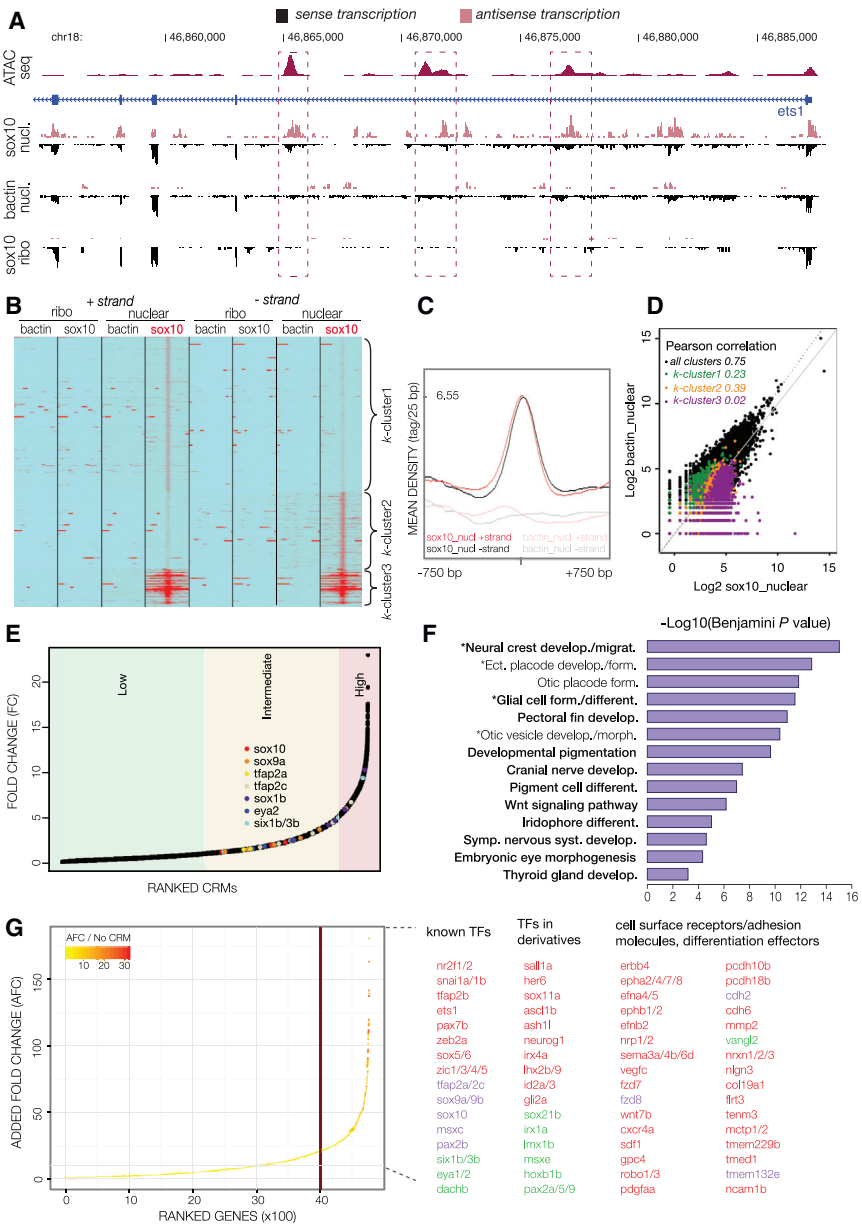


Figure 6. Analysis of Nuclear RNA Pools Reveals Bidirectional Transcription of Enhancers and Promoters as Unique Tissue-Specific Signature in the Nucleus

(A) Genome browser screenshot within *ets1* locus illustrating bidirectional transcription detected in NC nuclear but not ribosomal or *bactin* nuclear samples. Transcription within putative NC-specific CRMs is boxed in red.

(B) Heatmap depicting *k*-means linear enrichment clustering of strand-specific transcription across non-coding regions of open chromatin (ATAC peaks) in NC and *bactin* nuclear or ribosomal datasets.

(C) Mean density map of merged profiles for *k*-clusters 1 and 2 (from B) for *NC* and *bactin* transcripts.

(D) Scatterplot of transcriptional output between *NC* and *bactin* nuclear datasets for *k*-clusters: all ten clusters (black), clusters 1 (green), 2 (orange), and 3 (purple) (from B). Pearson correlation (*r*) is shown in the inset.

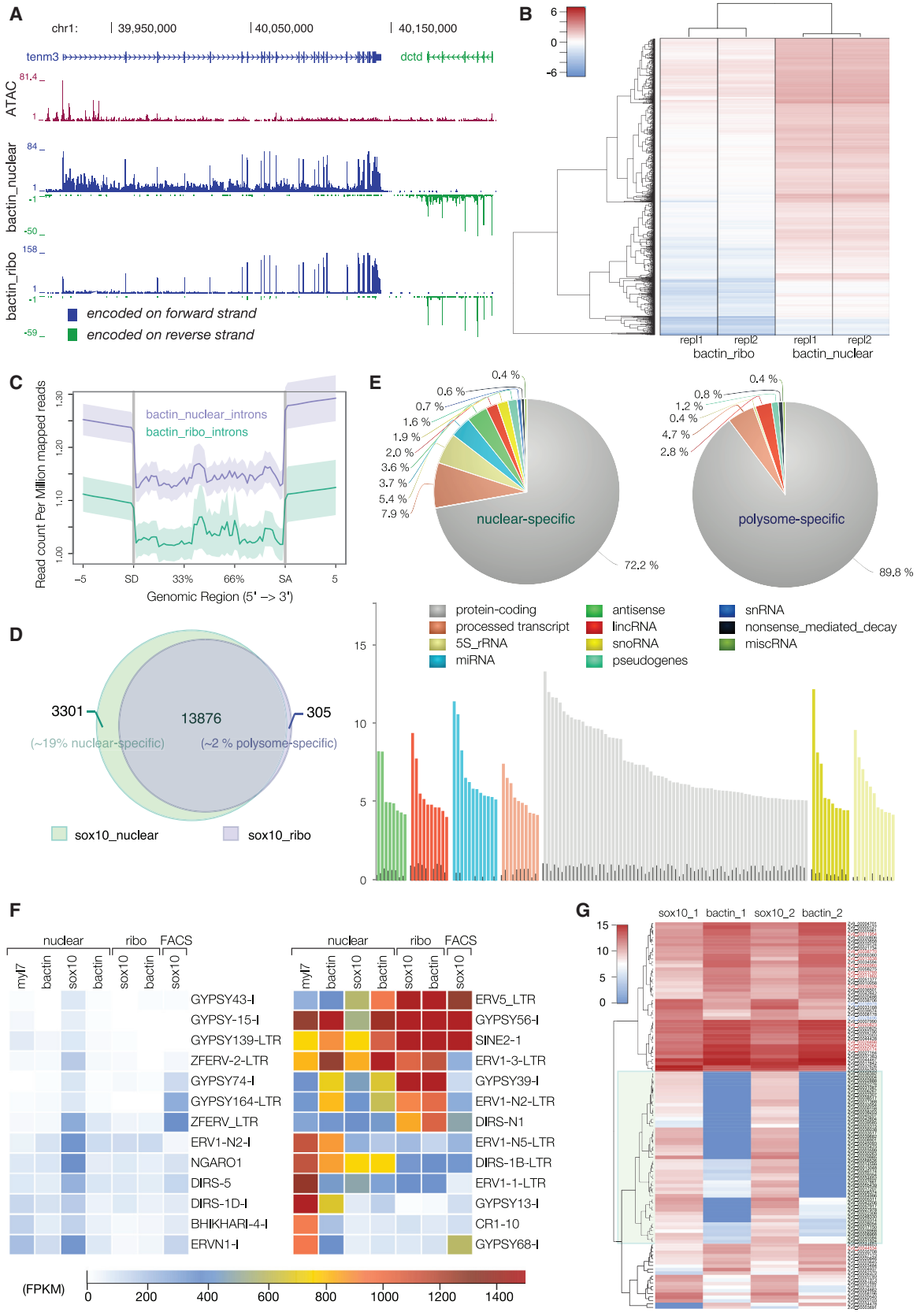
(E) Quantification of transcription from *k*-cluster 1 and 2 elements between NC and *bactin* nuclear dataset, ranked according to FC. Color dots represent annotated CRMs of known NC genes.

(F) GOs obtained by GREAT with transcribed elements for *k*-clusters 1 and 2.

(G) Genes ranked by AFC. Genes ranked beyond inflexion point listed with known involvement in NC development (red), otic placode/vesicle (green), or both (purple).

by the number of associated CRMs, we found that highly regulated loci, defined as those falling beyond the inflection point on the plot (Figure S6C), were controlled by at least three elements. A cumulative frequency graph showed that ~25% of loci were associated with three or more CRMs (Figure S6D).

The use of multiple enhancers to control the same locus may seem redundant, but their action on expression level is often thought to be additive (Arner et al., 2015). To uncover key NC regulators under the control of identified enhancers, we computed the additive fold change (AFC) as a sum of FCs of all active NC CRMs assigned to a given locus and ranked the loci according to their AFC value (total 4,767 genes). We then analyzed a set of highly regulated loci defined by AFC value falling beyond the inflection point (Figure 6G). These included genes coding for known TFs involved in specification of NC, ectodermal placodes, or both (Figure 6G) (Grocott et al., 2012; Simões-Costa and Bronner, 2015). In addition to a number of TFs involved in NC derivative fates (*ascl* and *ash* in sympathetic neurons, *neurogenin/lhx2b/her6/sox11a/irx4a* in sensory neurons, and *sall1a/irx1* in pectoral fin mesenchyme or *gli2a* in adrenal lineage), predominant categories include previously described signaling and cell-adhesion molecules involved in NC migration (e.g., *eph/ephrin*, *neuropilin/sema*, *wnt*, and *sdf1/cxcr4*). Given that the analyzed stage (16–18 ss) marks both migration and differentiation steps in NC ontogeny, a significant number of highly active loci encode for downstream effectors involved in terminal differentiation of NC derivatives. These include *neurexins* (NRXNs) and *neuroligins* (NLGNs), presynaptic cell-adhesion molecules secreted by sympathetic neurons, as well as *erbb4*, the neuregulin receptor involved in the differentiation of NC-derived glia. A significant number of highly regulated loci are transmembrane proteins (e.g., *tmed1*, *tmem229*, *bmctp1/2*, *flrt3*, *tmem132*, and *tenm3)*, consistent with the fact that NC cells rely heavily on cell-cell interactions with each other and their environment.



(legend on next page)

This analysis provides an insight into the migratory and differentiating NC regulatory programs, identifies a large number of NC regulatory factors, and provides a genome-wide representation of their upstream regulatory control. A full list of highly regulated NC loci is provided in Table S3.

Strand-Specific Profiling of NC RNA Landscapes in Different Subcellular Localizations

Biotagging enabled us to analyze transcriptional landscapes in different subcellular compartments within the same cell populations. The majority of gene expression studies use RNA from whole cells, overlooking the compartment-specific RNA composition, which is poised to reveal processes controlling expression, localization, and processing of RNA in the cell. Comparison of the *bactin* nuclear and *bactin* ribosomal datasets revealed significant differences in intronic RNA levels, consistent with the presence of immature transcripts in the nucleus and spliced mRNAs on ribosomes (Figure 7A). DESeq2 analysis, using introns of actively transcribed loci (ATAC_TSS set) as gene models, identified a group of coding genes with high intronic expression in nuclear, but not in ribosomal samples (Figure 7B). Merged intronic transcriptional profiles clearly showed this difference (Figure 7C). We characterized transcriptional patterns from different subcellular compartments at global scale (Figure S7A). The heatmaps obtained by clustering normalized datasets and visualizing them over coding regions indicated nearly identical merged profiles between replicates. However, we detected striking transcriptional pattern differences for nuclear and ribosomal samples, with the nuclear reads being maintained at similar levels over the entire gene body, characteristic of pervasive transcription across intronic regions. Profiles from FACS-purified whole NC cells, where the majority of transcripts (>90%) are cytosolic, and ribosomal NC samples were similar. They both feature a prominent central peak not seen in NC nuclear samples, which most likely corresponds to coding exons (Figure S7A). Such analyses demonstrate that our biotagging TRAP approach yields several-fold higher reads over coding loci compared to the biotagging INTACT approach (Figure S7A).

We compared expressed gene content in nuclei and on ribosomes in the NC cell population. Quantifying absolute gene activity (FPKM > 2) revealed a major overlap in transcribed gene content between the two subcellular compartments: ~20% of transcripts were found only in nuclei, and <2% were found only on ribosomes of NC cells (Figure 7D). The vast majority of transcripts correspond to protein-coding genes (72% in the nuclear pool and 90% in the ribosome pool). Further examinations uncovered a nuclear-specific demography consisting of regulatory RNAs that include small nucleolar RNA (snoRNAs), small nuclear RNA (snRNAs), primary microRNAs (miRNAs), 5 Svedberg units (s) rRNA, and antisense RNAs. In contrast, lncRNAs were equally represented in nuclei and ribosomes (Figure 7E). Unlike most cellular RNAs (Izaurralde and Mattaj, 1995), mature snoRNAs are not exported to the cytoplasm but remain to function in the nucleus (Terns et al., 1995). Thus, detection of snoRNAs in the nuclear samples further validates our approach. FPKM bar plots demonstrate clear differences in expression levels of representative nuclear compartment-specific RNA species (Figure 7E). Similar tendencies in subcellular-compartment-specific RNA content and diversity were observed in *bactin* subpopulation (Figures S7B and S7C). Thus, biotagging allows the investigation of gene expression at the tissue-specific and subcellular-compartment level.

Identification of Developmentally Regulated Non-coding RNAs that May Contribute to Tissue-Specific Gene Regulation

Since non-coding RNAs often overlap protein-coding regions on the opposite strand, strand-specific nuclear transcriptional profiling enables powerful analyses of the non-coding RNA landscape. More than 50% of the zebrafish genome sequence is seeded by type I and type II DNA transposable elements (TEs) (Howe et al., 2013). Although sometimes considered "junk" DNA, recent work suggests that TEs are involved in rewiring gene regulatory interactions during development (Gifford et al., 2013; Sundaram et al., 2014). Several studies have surveyed transcriptomes for TEs but often failed to recover tissue-specific TE transcription (Faulkner et al., 2009). A recent study using correlation of expression patterns across 18 different tissue types reveals systematic associations of particular TEs with certain tissues (Pavlicev et al., 2015).

We used our cardiomyocyte and NC datasets, along with ubiquitous controls at corresponding stages (16–18 ss and 26–30 hpf), to investigate whether TE expression is developmentally regulated. Differential expression analysis of all annotated classes of TEs in zebrafish across different datasets revealed that a number of TEs was expressed in a tissue-specific fashion and detected over a very broad spectrum of expression levels (Figure 7F). Several classes of differentially expressed TEs (i.e., *ERVN1-I*, *ERV1-N2-I*, *NGARO1*, and *ZFERV-2-LTR*) enriched in NC nuclei compared to the *bactin* samples were not found in ribosomal samples, suggesting that those elements are transcribed, but not exported. Given their relatively low expression

Figure 7. Comparative Genome-wide Profiles of Nuclear and Ribosomal Transcripts in the *sox10*-Positive Subpopulation Reveal Differences in Transcriptional Structure

(A) Genome browser screenshot illustrating intron retention in *bactin* nuclear, but not ribosomal, samples.

(B) Heatmap from differential expression analysis comparing intronic transcripts from *bactin* ribosomal and nuclear sample.

(C) Genome-wide additive expression profile of all differentially enriched introns larger than 30 kb, plotted based on intron read counts mapped per million reads.

(D) NC nuclear and ribosomal transcriptomes show significant overlap with 13,876 common annotated transcripts.

(E) Pie chart of nuclear- or ribosome-specific RNA species (top) and their respective FPKM values (bottom). The same color-code legend for RNA species was used in both the pie chart and the bar plot. In the bar plot, color and black bars correspond to FPKM values from nuclear and ribosomal samples, respectively.

(F) Heatmap of TEs expressed in a tissue-specific and dynamic fashion across different samples. Low-level-expressing enhancer TEs (left) and high-level-expressing exported TEs (right) are shown.

(G) Heatmap of differentially expressed lncRNAs in NC versus *bactin* nuclei. Green frame, NC-specific lncRNAs expressed at negligent levels in the *bactin* sample.

levels, such NC-specific TEs may primarily function as enhancers (Pavlicev et al., 2015). In contrast, we uncovered a set of TEs that were transcribed at very high levels in NC but detected mostly in the ribosomal compartment, suggesting these rapidly exported TEs are likely contained within mature coding transcripts and unlikely to act in *cis* (e.g., *ERV1-N2-LTR*, *DIRS-N1*, and *GYPSY39-I*). We identified a group of elements (*CR1-10*, *GYPSY13-I*, *GYPSY68-I*, and *ERV1-1-LTR*) transcribed at high levels specifically in cardiomyocytes. Thus, TE expression appears to be developmentally regulated in both a cell-type and subcellular-compartment manner.

Landmark studies that identified and characterized lncRNAs using large-scale transcriptomics and histone chromatin immunoprecipitation sequencing (ChIP-seq) have thrust these molecules into the spotlight as potential fine-tuners of gene expression (Guttman et al., 2009) by forming molecular scaffolds to recruit chromatin regulators (Wang et al., 2011). Some recent reports suggest that lncRNA production, rather than the lncRNA transcripts themselves, influences gene expression of neighboring genes in *cis* (Engreitz et al., 2016). Additional studies attempting to dissect the biology of lncRNAs have highlighted the importance of cellular compartmentalization. While lncRNAs were initially described as present in nuclei (Derrien et al., 2012), the use of ribosome footprinting in genome-wide studies made it evident that lncRNAs can associate with ribosomes (Guttman et al., 2013; Ingolia et al., 2014). By showing that transcripts associated with ribosomes may not be translated into proteins but could be regulating or be regulated by the process of translation, such findings have challenged the central dogma of translation on ribosomes as a one-way process.

Our biotagging approach is well suited to exploring questions involving lncRNA function and localization. As proof of concept, we have quantified known zebrafish lncRNAs (Pauli et al., 2012), identifying lncRNAs that were differentially regulated between cell types and compartments. We identified 51 differentially expressed lncRNAs ($p < 0.05$) in the *myl7* versus *bactin* nuclei (26–30 hpf; Figure S7E) and 111 lncRNAs differentially expressed in *sox10* nuclei (versus *bactin* nuclei; 16–18 ss; Figure 7G). Only three lncRNAs were detected when comparing *sox10* versus *bactin* ribosomal pools (16-18ss; data not shown). NC- and myocardial-specific lncRNA sets contain 14 common lncRNAs (Figure S7F), but these mostly represent highly expressed species found in NC and whole embryos at earlier stages (16–18 ss) that are downregulated in differentiating myocardium at 26–30 hpf. The majority of unique non-overlapping NC-specific lncRNAs represent highly expressed specifically enriched species (Figure 7G, framed). Our results on migrating NC show that lncRNAs can be found on ribosomes as described previously, but developmentally regulated lncRNAs are more likely to be enriched in nuclei. Therefore, our biotagging approach in zebrafish offers a better means to identify cell-type-specific lncRNAs and provides the subcellular resolution required for studies of their biological function in development.

Conclusions

Deciphering the intricacies of developmental programs in specific cell types requires the ability to isolate defined, small subpopulations from their in vivo context. Our binary genetic toolkit enables in vivo biotinylation of proteins in defined compartments (nuclei and ribosomes) and cell populations of interest, permitting the isolation of biotinylated proteins and their interacting molecular components with high stringency. Although typical genome-wide assays require large amounts of starting material, the stringency, negligible background, and minimal variability of the biotagging toolkit enable us to robustly identify even unstable RNA species from specific cell subpopulations of the developing embryo (cardiomyocytes, ~400 cells per embryo; NC, ~2,000 cells per embryo). No complex amplification schemes were required for transcriptome profiling. The biotagging toolkit presented here, containing seven tissue-specific and four ubiquitous BirA driver lines, as well as the five ubiquitous Avi-effectors (see Table S1), can easily be expanded using BirA constructs featuring BirA open reading frame (ORF) donors for generation of new drivers by either BAC recombineering or conventional plasmid transgenesis. Together, the toolkit enables epigenomic, transcriptional, and proteomic profiling of individual cell types within the heterogeneous context of developing embryos or zebrafish models of human disease.

As the versatility and modularity of the biotagging toolkit allows the rapid isolation of RNA species from compartments of specific cell types, we used it to characterize nuclear and ribosomal transcriptomes from migrating NC cells and differentiating cardiomyocytes at different stages of development. At 16–18 ss, genome-wide chromatin accessibility assays show that the nuclei of both the NC and the majority of the early embryo present a broad open chromatin architecture, resulting in pervasive divergent transcription. We find that this phenomenon is more prominent in NC than in whole-embryo nuclei at early stages of development, consistent with their stem cell-like nature. Canonical differential expression analyses across coding loci of total nuclear transcriptomes in NC versus whole embryo did not recover a clear NC transcriptional profile, further supporting this idea. Similar analysis of *myl7* nuclear samples at the later developmental stage (26–30 hpf) clearly recovered the cardiomyocyte transcriptional signature.

Interestingly, we discovered that tissue-specific gene regulatory logic is encrypted in nuclear transcriptomes primarily at the level of CRMs (enhancers) and other non-coding species (lncRNAs and transposons). By quantifying bidirectional transcription of enhancers, detected specifically in NC, but not in whole-embryo nuclei, we uncovered the ensemble of putative CRMs controlling NC identity at 16–18 ss. Thus, using the biotagging approach, we gained a holistic insight into the regulatory landscape and transcriptional signature of migrating NC cells. This study highlights how a cohort of non-coding elements expressed in the nucleus modulates NC gene regulatory program, demonstrating that more than the transcription of protein-coding genes shapes the migratory NC identity.

EXPERIMENTAL PROCEDURES

Zebrafish Husbandry

This study was carried out in accordance to procedures authorized by the UK Home Office in accordance with UK law (Animals [Scientific Procedures] Act 1986) and the recommendations in the *Guide for the Care and Use of Laboratory Animals*. Adult fish were maintained as described previously (Westerfield, 2000).

Generation of Biotagging Toolkit

Constructs (plasmid and BAC) for generating biotagging transgenic lines (now available from Addgene) were co-injected with tol2 mRNA into one-cell-stage zebrafish embryos. Injected F_0s were raised and screened for founders. Positive F1s grown to reproductive age were crossed for biotagging experiments.

Nuclei and Polysomal Isolation

100–350 embryos per experiment were washed and lysed in hypotonic buffer for nuclei isolation or optimized Cell Lysis Buffer for polysomal isolation using a Dounce homogenizer. Lysates were cleared by centrifugation, and nuclei or polysome pellets were washed using buffers specifically adapted to each procedure and incubated with Streptavidin magnetic beads. The bead-nuclei or bead-polysome complexes were captured using a flow-based setup (nuclei) or magnetic separation setup (polysomes) and lysed for total RNA extractions.

Library Preparation and Next Generation Sequencing

Non-directional sequencing libraries were built using NEBNext Ultra RNA library kit for Illumina (New England Biolabs) starting from poly(A)-selected RNA transcripts. Directional RNA-sequencing libraries were prepared using Stranded RNA-Seq Library Preparation Kit (KAPA Biosystems) starting from ribo-depleted total nuclear RNA or polysomal RNA. Next generation sequencing (NGS) was performed on HiSeq2500 or Nextseq500 Illumina platforms.

Bioinformatics Processing

ATAC-Seq

Trimmed reads were mapped using bowtie (v.1.0.0) as described previously (Buenrostro et al., 2013). Peak calling was performed using MACS2 with –nomodel and –slocal 1,000 parameters (Zhang et al., 2008).

RNA-Seq

After mapping, compressed binary version of the sequence alignment/map (BAM) files were split according to strand using custom scripts available at https://github.com/tsslab/biotagging/. Differential expression analyses were performed using DESeq2 (coding genes, introns, and lncRNAs) (Love et al., 2014) and using the rank product non-parametric method (TEs) (Göke et al., 2015). GSEAs were performed using the Piano package (Väremo et al., 2013) and functional classifications using the Panther system (http://www.pantherdb.org/). Statistical overrepresentation was calculated using hypergeometric and exact Fisher's tests (Mi et al., 2013).

k-means Clustering

k-means clustering of ATAC_TSS or ATAC_enhancer elements based on *sox10* nuclear, *bactin* nuclear, and *sox10* polysomal strand-specific RNA-seq patterns was performed using the seqMINER platform (Ye et al., 2011).

Ranking NC-Specific CRMs

NC-specific CRMs were ranked according to their FC value, computed as the ratio of FPKM expression value in *sox10* and *bactin* sample for each ATAC_enhancer feature. These CRMs were assigned to the proximal expressed gene targets using bedtools and GREAT (McLean et al., 2010). AFC, used to quantify the effect of multiple enhancers, was computed as a sum of FCs of all active CRMs assigned to a given locus. Functional analysis of identified CRMs was performed using GREAT and statistical significance computed by both binominal and hypergeometric tests (McLean et al., 2010).

A detailed description of the toolkit construction and validation, optimized isolation protocols with specific buffer compositions and parameters for bioinformatics processing, including *k*-means analyses, are available in Supplemental Experimental Procedures.

ACCESSION NUMBERS

The accession number for all the NGS data reported in this study is GEO: GSE89670 and are available via https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE89670. The data are also available via the DANIO-CODE consortium (https://danio-code.zfin.org/daniocode/). Custom scripts associated with this study are available at https://github.com/tsslab/biotagging/.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, seven figures, and three tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2017.03.045.

AUTHOR CONTRIBUTIONS

Conceptualization, L.A.T., S.E.F., and T.S.-S.; Methodology, L.A.T., V.C.-M., T.H.-H., U.S., and T.S.-S.; Validation, V.C.-M., L.A.T., and D.G.; Investigation, V.C.-M., L.A.T., D.G., and T.S.-S.; Writing – Original Draft, L.A.T. and T.S.-S.; Writing – Review & Editing, L.A.T., V.C.-M., D.G., S.E.F., and T.S.-S.; Funding Acquisition, S.E.F. and T.S.-S.; Resources, L.A.T. and V.C.-M.; Data Curation, D.G.; Supervision, T.S.-S.

ACKNOWLEDGMENTS

This work was supported by a March of Dimes Basil O'Connor Award (#5-FY12-564), a Lister Institute Research Prize, and an Oxford BHF CRE award (#RE/08/004, to T.S.-S.), a Clarendon Fund Fellowship (to V.C.-M.), and an SNF Fellowship (PBSKP3_145791, to D.G.). We thank Tudor Fulga, Ferdinand Marlétaz, and Simon Restrepo for comments on the manuscript.

Received: August 1, 2014
Revised: December 21, 2016
Accepted: March 13, 2017
Published: April 11, 2017

REFERENCES

Amin, N.M., Greco, T.M., Kuchenbrod, L.M., Rigney, M.M., Chung, M.I., Wallingford, J.B., Cristea, I.M., and Conlon, F.L. (2014). Proteomic profiling of cardiac tissue by isolation of nuclei tagged in specific cell types (INTACT). Development *141*, 962–973.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al.; FANTOM Consortium (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455–461.

Andersson, R., Sandelin, A., and Danko, C.G. (2015). A unified architecture of transcriptional regulatory elements. Trends Genet. *31*, 426–433.

Arner, E., Daub, C.O., Vitting-Seerup, K., Andersson, R., Lilje, B., Drabløs, F., Lennartsson, A., Rönnerblad, M., Hrydziuszko, O., Vitezic, M., et al.; FANTOM Consortium (2015). Transcribed enhancers lead waves of coordinated transcription in transitioning mammalian cells. Science *347*, 1010–1014.

Barthelson, R.A., Lambert, G.M., Vanier, C., Lynch, R.M., and Galbraith, D.W. (2007). Comparison of the contributions of the nuclear and cytoplasmic compartments to global gene expression in human cells. BMC Genomics *8*, 340.

Boettiger, A.N., and Levine, M. (2009). Synchronous and stochastic patterns of gene activation in the Drosophila embryo. Science *325*, 471–473.

Brade, T., Männer, J., and Kühl, M. (2006). The role of Wnt signalling in cardiac development and tissue remodelling in the mature heart. Cardiovasc. Res. *72*, 198–209.

Bradford, Y., Conlin, T., Dunn, N., Fashena, D., Frazer, K., Howe, D.G., Knight, J., Mani, P., Martin, R., Moxon, S.A., et al. (2011). ZFIN: enhancements and updates to the Zebrafish Model Organism Database. Nucleic Acids Res. *39*, D822–D829.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods *10*, 1213–1218.

Core, L.J., Waterfall, J.J., and Lis, J.T. (2008). Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science *322*, 1845–1848.

Core, L.J., Martins, A.L., Danko, C.G., Waters, C.T., Siepel, A., and Lis, J.T. (2014). Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. Nat. Genet. *46*, 1311–1320.

Cronan, J.E., Jr. (1990). Biotination of proteins in vivo. A post-translational modification to label, purify, and study proteins. J. Biol. Chem. *265*, 10327–10333.

de Boer, E., Rodriguez, P., Bonte, E., Krijgsveld, J., Katsantoni, E., Heck, A., Grosveld, F., and Strouboulis, J. (2003). Efficient biotinylation and single-step purification of tagged transcription factors in mammalian cells and transgenic mice. Proc. Natl. Acad. Sci. USA *100*, 7480–7485.

De Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B.K., Muller, H., Ragoussis, J., Wei, C.L., and Natoli, G. (2010). A large fraction of extragenic RNA pol II transcription sites overlap enhancers. PLoS Biol. *8*, e1000384.

Deal, R.B., and Henikoff, S. (2010). A simple method for gene expression and chromatin profiling of individual cell types within a tissue. Dev. Cell *18*, 1030–1040.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., et al. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res. *22*, 1775–1789.

Engreitz, J.M., Haines, J.E., Perez, E.M., Munson, G., Chen, J., Kane, M., McDonel, P.E., Guttman, M., and Lander, E.S. (2016). Local regulation of gene expression by lncRNA promoters, transcription and splicing. Nature *539*, 452–455.

Faulkner, G.J., Kimura, Y., Daub, C.O., Wani, S., Plessy, C., Irvine, K.M., Schroder, K., Cloonan, N., Steptoe, A.L., Lassmann, T., et al. (2009). The regulated retrotransposon transcriptome of mammalian cells. Nat. Genet. *41*, 563–571.

Gaertner, B., Johnston, J., Chen, K., Wallaschek, N., Paulson, A., Garruss, A.S., Gaudenz, K., De Kumar, B., Krumlauf, R., and Zeitlinger, J. (2012). Poised RNA polymerase II changes over developmental time and prepares genes for future expression. Cell Rep. *2*, 1670–1683.

Gessert, S., and Kühl, M. (2010). The multiple phases and faces of wnt signaling during cardiac differentiation and development. Circ. Res. *107*, 186–199.

Gifford, W.D., Pfaff, S.L., and Macfarlan, T.S. (2013). Transposable elements as genetic regulatory substrates in early development. Trends Cell Biol. *23*, 218–226.

Göke, J., Lu, X., Chan, Y.S., Ng, H.H., Ly, L.H., Sachs, F., and Szczerbinska, I. (2015). Dynamic transcription of distinct classes of endogenous retroviral elements marks specific populations of early human embryonic cells. Cell Stem Cell *16*, 135–141.

Grocott, T., Tambalo, M., and Streit, A. (2012). The peripheral sensory nervous system in the vertebrate head: a gene regulatory perspective. Dev. Biol. *370*, 3–23.

Guenther, M.G., Levine, S.S., Boyer, L.A., Jaenisch, R., and Young, R.A. (2007). A chromatin landmark and transcription initiation at most promoters in human cells. Cell *130*, 77–88.

Guttman, M., Amit, I., Garber, M., French, C., Lin, M.F., Feldser, D., Huarte, M., Zuk, O., Carey, B.W., Cassady, J.P., et al. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature *458*, 223–227.

Guttman, M., Russell, P., Ingolia, N.T., Weissman, J.S., and Lander, E.S. (2013). Ribosome profiling provides evidence that large noncoding RNAs do not encode proteins. Cell *154*, 240–251.

Heiman, M., Schaefer, A., Gong, S., Peterson, J.D., Day, M., Ramsey, K.E., Suárez-Fariñas, M., Schwarz, C., Stephan, D.A., Surmeier, D.J., et al. (2008). A translational profiling approach for the molecular characterization of CNS cell types. Cell *135*, 738–748.

Hendrix, D.A., Hong, J.W., Zeitlinger, J., Rokhsar, D.S., and Levine, M.S. (2008). Promoter elements associated with RNA Pol II stalling in the Drosophila embryo. Proc. Natl. Acad. Sci. USA *105*, 7762–7767.

Housley, M.P., Reischauer, S., Dieu, M., Raes, M., Stainier, D.Y., and Vanhollebeke, B. (2014). Translational profiling through biotinylation of tagged ribosomes in zebrafish. Development *141*, 3988–3993.

Howe, K., Clark, M.D., Torroja, C.F., Torrance, J., Berthelot, C., Muffato, M., Collins, J.E., Humphray, S., McLaren, K., Matthews, L., et al. (2013). The zebrafish reference genome sequence and its relationship to the human genome. Nature *496*, 498–503.

Ingolia, N.T., Brar, G.A., Stern-Ginossar, N., Harris, M.S., Talhouarne, G.J., Jackson, S.E., Wills, M.R., and Weissman, J.S. (2014). Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. Cell Rep. *8*, 1365–1379.

Izaurralde, E., and Mattaj, I.W. (1995). RNA export. Cell *81*, 153–159.

Kim, T.K., Hemberg, M., Gray, J.M., Costa, A.M., Bear, D.M., Wu, J., Harmin, D.A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., et al. (2010). Widespread transcription at neuronal activity-regulated enhancers. Nature *465*, 182–187.

Kowalczyk, M.S., Hughes, J.R., Garrick, D., Lynch, M.D., Sharpe, J.A., Sloane-Stanley, J.A., McGowan, S.J., De Gobbi, M., Hosseini, M., Vernimmen, D., et al. (2012). Intragenic enhancers act as alternative promoters. Mol. Cell *45*, 447–458.

Lepoivre, C., Belhocine, M., Bergon, A., Griffon, A., Yammine, M., Vanhille, L., Zacarias-Cabeza, J., Garibal, M.A., Koch, F., Maqbool, M.A., et al. (2013). Divergent transcription is associated with promoters of transcriptional regulators. BMC Genomics *14*, 914.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol. *28*, 495–501.

Mi, H., Muruganujan, A., Casagrande, J.T., and Thomas, P.D. (2013). Large-scale gene function analysis with the PANTHER classification system. Nat. Protoc. *8*, 1551–1566.

Mitchell, J.A., Clay, I., Umlauf, D., Chen, C.Y., Moir, C.A., Eskiw, C.H., Schoenfelder, S., Chakalova, L., Nagano, T., and Fraser, P. (2012). Nuclear RNA sequencing of the mouse erythroid cell transcriptome. PLoS ONE *7*, e49274.

Nepal, C., Hadzhiev, Y., Previti, C., Haberle, V., Li, N., Takahashi, H., Suzuki, A.M., Sheng, Y., Abdelhamid, R.F., Anand, S., et al. (2013). Dynamic regulation of the transcription initiation landscape at single nucleotide resolution during vertebrate embryogenesis. Genome Res. *23*, 1938–1950.

Ooi, S.L., Henikoff, J.G., and Henikoff, S. (2010). A native chromatin purification system for epigenomic profiling in Caenorhabditis elegans. Nucleic Acids Res. *38*, e26.

Pauli, A., Valen, E., Lin, M.F., Garber, M., Vastenhouw, N.L., Levin, J.Z., Fan, L., Sandelin, A., Rinn, J.L., Regev, A., and Schier, A.F. (2012). Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. Genome Res. *22*, 577–591.

Pavlicev, M., Hiratsuka, K., Swaggart, K.A., Dunn, C., and Muglia, L. (2015). Detecting endogenous retrovirus-driven tissue-specific gene transcription. Genome Biol. Evol. *7*, 1082–1097.

Preker, P., Nielsen, J., Kammler, S., Lykke-Andersen, S., Christensen, M.S., Mapendano, C.K., Schierup, M.H., and Jensen, T.H. (2008). RNA exosome depletion reveals transcription upstream of active human promoters. Science *322*, 1851–1854.

Seila, A.C., Calabrese, J.M., Levine, S.S., Yeo, G.W., Rahl, P.B., Flynn, R.A., Young, R.A., and Sharp, P.A. (2008). Divergent transcription from active promoters. Science *322*, 1849–1851.

Sigova, A.A., Mullen, A.C., Molinie, B., Gupta, S., Orlando, D.A., Guenther, M.G., Almada, A.E., Lin, C., Sharp, P.A., Giallourakis, C.C., and Young, R.A. (2013). Divergent transcription of long noncoding RNA/mRNA gene pairs in embryonic stem cells. Proc. Natl. Acad. Sci. USA *110*, 2876–2881.

Simões-Costa, M., and Bronner, M.E. (2015). Establishing neural crest identity: a gene regulatory recipe. Development *142*, 242–257.

Steiner, F.A., Talbert, P.B., Kasinathan, S., Deal, R.B., and Henikoff, S. (2012). Cell-type-specific nuclei purification from whole animals for genome-wide expression and chromatin profiling. Genome Res. *22*, 766–777.

Sundaram, V., Cheng, Y., Ma, Z., Li, D., Xing, X., Edge, P., Snyder, M.P., and Wang, T. (2014). Widespread contribution of transposable elements to the innovation of gene regulatory networks. Genome Res. *24*, 1963–1976.

Terns, M.P., Grimm, C., Lund, E., and Dahlberg, J.E. (1995). A common maturation pathway for small nucleolar RNAs. EMBO J. *14*, 4860–4871.

Tryon, R.C., Pisat, N., Johnson, S.L., and Dougherty, J.D. (2013). Development of translating ribosome affinity purification for zebrafish. Genesis *51*, 187–192.

Väremo, L., Nielsen, J., and Nookaew, I. (2013). Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. Nucleic Acids Res. *41*, 4378–4391.

Wang, K.C., Yang, Y.W., Liu, B., Sanyal, A., Corces-Zimmerman, R., Chen, Y., Lajoie, B.R., Protacio, A., Flynn, R.A., Gupta, R.A., et al. (2011). A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. Nature *472*, 120–124.

Westerfield, M. (2000). The Zebrafish Book. A Guide for the Laboratory Use of Zebrafish (Danio rerio), Fourth Edition (University of Oregon Press).

Ye, T., Krebs, A.R., Choukrallah, M.A., Keime, C., Plewniak, F., Davidson, I., and Tora, L. (2011). seqMINER: an integrated ChIP-seq data interpretation platform. Nucleic Acids Res. *39*, e35.

Zabidi, M.A., Arnold, C.D., Schernhuber, K., Pagani, M., Rath, M., Frank, O., and Stark, A. (2015). Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. Nature *518*, 556–559.

Zaghlool, A., Ameur, A., Nyberg, L., Halvardson, J., Grabherr, M., Cavelier, L., and Feuk, L. (2013). Efficient cellular fractionation improves RNA sequencing analysis of mature and nascent transcripts from human tissues. BMC Biotechnol. *13*, 99.

Zeitlinger, J., Stark, A., Kellis, M., Hong, J.W., Nechaev, S., Adelman, K., Levine, M., and Young, R.A. (2007). RNA polymerase stalling at developmental control genes in the Drosophila melanogaster embryo. Nat. Genet. *39*, 1512–1516.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., and Liu, X.S. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.