**Cell Reports**

**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 © 2017 The Authors.

http://dx.doi.org/10.1016/j.celrep.2017.03.045
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, *

1Molecular and Computational Biology, University of Southern California, Los Angeles, CA 90089, USA
2Radcliffe Department of Medicine, Weatherall Institute of Molecular Medicine, University of Oxford, Oxford OX3 9DS, UK
3Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91125, USA
4Co-first author
5Lead 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 sub-populations 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...
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) recombinering, 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 (Hei- man 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 (jactin) or ubiquitin (ubiq) promoters to drive ubiquitous expression of a zebrafish-compatible Avi-Cerulean-RanGap or Avi-Cereulan-Rpi10 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(jactin) 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...
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 \( \frac{C}{C_{24}} \% \) 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(\( iactin \)) drivers with the nucAvi(\( iactin \)) 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,
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 nuclear datasets at 2 FPKMs or higher. A statistically significant number of them (133/419, p < 0.01) were overrepresented in nuclear datasets versus control embryos. Furthermore, a statistically significant number of myocardial genes (Figure S3C). Biotyping 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 this 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(jactin) 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; Sela 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 (Cl1-5, 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
downstream coding genes and may represent the earliest event in the gene activation cascade (Amer 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 (RC11 = 0.23, RC12 = 0.39, and RC33 = 0.02; Figure 6D) are significantly offset from the coefficient for all clusters (RC11 = 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 1A 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...
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; Simoes-Costa and Bronner, 2015). In addition to a number of TFs involved in NC derivative fates (ascl1 and ash in sympathetic neurons, neurogenin/lhx2b/her6/sox11a/irx4a in sensory neurons, and sal1a/irx1 in pectoral fin mesenchyme or glio2a 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 sal1a/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 erb4, the neuregulin receptor involved in the differentiation of NC-derived glia. A significant number of highly regulated loci are transmembrane proteins (e.g., tmem1, tmem229, bmctp1/2, fnt3, tmem132, and tenm3), consistent with the fact that NC cells rely heavily on cell-cell interactions with each other and their environment.
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). Analyses such as demonstrated 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 nuclear RNA (snRNAs), small nuclear RNA (snRNAs), primary microRNAs (miRNAs), 5 Svedberg units (s) rRNA, and antisense RNAs. In contrast, IncRNAs were equally represented in nuclei and ribosomes (Figure 7E). Unlike most cellular RNAs (Izaurralde and Mattaj, 1995), mature snRNAs 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., ERV1-N1-I, ERV1-N2-I, NARG01, 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...
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, DIRSN1, 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 IncRNAs 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 IncRNA production, rather than the IncRNA transcripts themselves, influences gene expression of neighboring genes in cis (Engreitz et al., 2016). Additional studies attempting to dissect the biology of IncRNAs have highlighted the importance of cellular compartmentalization. While IncRNAs were initially described as present in nuclei (Derrien et al., 2012), the use of ribosome footprinting in genome-wide studies made it evident that IncRNAs 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 IncRNA function and localization. As proof of concept, we have quantified known zebrafish IncRNAs (Pauli et al., 2012), identifying IncRNAs that were differentially regulated between cell types and compartments. We identified 51 differentially expressed IncRNAs (p < 0.05) in the myocardium at 26–30 hpf. The majority of unique non-overlapping IncRNAs (Figure S7F), but these mostly represent highly expressed species found in NC and whole embryos at earlier stages (16–18 ss; Figure 7G). Only three IncRNAs were detected when comparing sox10 versus bactin ribosomal pools (16-18ss; data not shown). NC- and myocardial-specific IncRNA sets contain 14 common IncRNAs (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 IncRNAs represent highly expressed specifically enriched species (Figure 7G, framed). Our results on migrating NC show that IncRNAs can be found on ribosomes as described previously, but developmentally regulated IncRNAs are more likely to be enriched in nuclei. Therefore, our biotagging approach in zebrafish offers a better means to identify cell-type-specific IncRNAs 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 recombinering 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 (IncRNAs 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).
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 (PBSPK3_145791, to D.G.). We thank Tudor Fulga, Ferdinand Marletz, 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.J., 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., Schmid, 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 Danilo, C.G. (2015). A unified architecture of transcriptional regulatory elements. Trends Genet. 31, 426–433.

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

Bartholmai, 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., Mannan, J., and Kuhl, 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.

Buonrostro, 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., Krigsveld, J., Katsantonis, 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., Tusì, 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., Gutmann, M., and Lander, E.S. (2016). Local regulation of gene expression by IncRNA promoters, transcription and splicing. Nature 539, 452–455.

Faulkner, G.J., Kimura, Y., Daub, C.O., Wani, S., Plessy, C., Irvine, K.M., 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.

Housley, M.P., Reischauer, S., Dieu, M., Raes, M., Stainier, D.Y., and Vanhollee, 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., Berthalot, 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., Stem-Ginossar, N., Harris, M.S., Talhouarne, G.J., Jackson, S.E., Willis, M.R., and Weissman, J.S. (2014). Ribosome profiling reveals pervasive translation outside of annotated protein-coding genes. Cell 168, 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-Hailey, 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., Vennimmen, 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., Hillier, M., Clarke, S.L., Schara, 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., Eskivi, 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.

Paul, 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 Mugila, L. (2015). Detecting endogenous retrovirus-driven tissue-specific gene transcription. Genome Biol. 17, 1082–1097.

Preker, P., Nielsen, J., Kammiller, 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.

Pellisé-Cabrera, 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., Grubb, 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., Kelis, M., Hong, J.W., Nечаев, 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.
Supplemental Information

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

Le A. Trinh, Vanessa Chong-Morrison, Daria Gavriouchkina, Tatiana Hochgreb-Hägele, Upeka Senanayake, Scott E. Fraser, and Tatjana Sauka-Spengler
Figure S1. Biotagging *sox10* BirA drivers and Avi-tagged RanGap effectors. Related to Figures 1 and 2.
A-C Differential expression between \textit{sox10} biotagging transgenic and BAC drivers. Wide-field image of $\text{TgBAC(Sox10:BirA-mCherry)}^{\text{ox104a}}$ (ncBirA(BAC)) (A). $\text{Tg(Sox10:BirA-membCherry)}^{\text{ct706a}}$ (ncBirA) and $\text{Tg(Sox10:BirA-membCherry)}^{\text{ct706b}}$ (ncBirA(b)) transgenes at 24hpf (B, C). The two alleles (B, C) of \textit{sox10} Biotagging transgenic exhibit expression in different neural crest derivatives that are included in the overall \textit{sox10} expression pattern, by \textit{sox10} Biotagging BAC (A). Arrow points to lack of expression in the otic vesicle, while arrowheads point to lack of expression in the midbrain of \textit{sox10} transgenes.

D-L Biotagging Avi-tagged nuclear localized effectors. Schematic of two variant Avi-tagged RanGap constructs for generating Avi effector transgenes. N-terminal Avi-tag construct (Avi-Cerulean-Rangap) contains the beta-actin2 ($\beta$actin) promoter upstream of Avi-tag (steelblue), the Tobacco Etch Virus protease cleave site (TeV, green), Cerulean (turquoise), the C-terminal domain of RanGap (purple), and a polyA signal (D). C-terminal Avi-tag construct (RanGap-Cerulean-Avi) contains the beta-actin2 ($\beta$actin) promoter upstream of the C-terminal domain of RanGap (purple), Cerulean (turquoise), the Tobacco Etch Virus protease cleave site (TeV, green), Avi-tag (steelblue), and a polyA signal (E). Both constructs are flanked by tol2 elements (yellow) for transgenesis by Tol2 transposition. 3-D projection of confocal Z-stack of Avi-RanGap (F-G) and RanGap-Avi (H-I) of the developing inner ear (F, H) and somite (G, I), imaged at 32hpf. Both Avi-Cerulean-RanGap and RanGp-Cerulean-Avi proteins localize similarly to the nucleus of all cells in the embryo. Schematic of Avi-RanGap effector construct with ubiquitin promoter ($\text{ubiQ}$) upstream of N-terminal Avi tagged RanGap elements (J). Confocal image of hindbrain (K) and eye (L) of $\text{Tg(ubiQ:Avi-RanGap)}$ embryo (nucAvi($\text{ubiQ}$)). Scale bars: 20\,\mu m, except 50\,\mu m in (K, L).
Figure S2. Biotinylation of Avi-tagged protein is dependent on the level of expression. Related to Figure 3.

A Streptavidin Western blot of nuclear extracts from embryos injected with Avi-RanGap (lane 1), BirA (lane 2), both mRNA (lane 3) and wildtype un-injected embryos (lane 4). Arrow points to biotinylated Avi-RanGap that is specifically labelled with Streptavidin conjugate in lysates from embryos co-injected with Avi-RanGap and BirA (lane 3). Asterisks indicate endogenously biotinylated proteins also present in un.injected wildtype embryos. BirA expression (lane 2) does not elevate endogenous biotinylation over background level (lane 4). Avi-tag is insensitive to biotinylation by endogenous biotin ligases (line 1 compared to line 3).

B Streptavidin (upper panel), Anti-GFP (middle panel) and Anti-HA (bottom panel) Western blot of whole cell extract from wildtype (lane 1); TgBAC(sox10:BirA) (ncBirA(BAC)) (lane 3); Tg(sox10:BirA) (ncBirA) (lane 4); Tg(ubiq:AviRpl10) (riboAvi(ubiq)) (lane 2); or double transgenic of Avi-tag and BirA (lanes 5, 6, 7, 8) embryos. Number of plus marks indicate level of expression described to the right of blots. Avi-Rpl10 is biotinylated only when highly expressed (lanes 6, 7, 8). Anti-GFP western blot detects a cleaved Avi-Rpl10 that is 30 kDa (red box in upper and middle panels). Anti-HA Western blot detects BirA that contains HA-tag (red box in bottom panel).

C Streptavidin (upper left panel), Anti-GAPDH (lower left panel) and Anti-GFP (right panel) Western blot of whole cell extract from wildtype (lane 1); TgBAC(sox10:BirA) (lane 4); or double transgenic of Avi-Rpl10 and BirA (lanes 2 and 3) embryos in which BirA expression varies while “Avi-Rp10” expression is high. Anti-GFP Western blot detects both cleaved and uncleaved Avi-Rpl10 that is 30 kDa and 55 kDa, respectively.

Figure legend (B and C)

For driver (sox10:BirA) and effector (Avi-Rpl10) alleles:

*+: Protein expression only, no fluorescence detected at 24hpf
++*: Protein expression, AND fluorescence detected at 24hpf
+: No protein expression and no fluorescence detected at 24hpf
Figure S3. Enrichment of cell-type specific gene expression in developing cardiomyocytes and pairwise comparison of biological duplicates. Related to Figure 4.

A, B Barplot of FPKM values from myl7 nuclear datasets for genes annotated in ZFIN expression database as detected in myocardium from 0-24hpf. C Quantification of relative enrichment of myocardial transcripts as determined by RT-qPCR using cDNA isolated from nuclei purified from embryos double transgenic for Tg(myl7:BirA-membCherry)\textsuperscript{770ha}, Tg(bactin:Avi-Cerulean-RanGap)\textsuperscript{770sb} (myoBirA;nuClAvi(bact)) compared to cDNA extracted from whole embryos. Relative levels of transcripts for myosin light polypeptide 7 (myl7), ventricular myosin heavy chain (vmhc) and SLU7 splicing factor homolog (slu7) were normalized to glyceraldehyde 3-phosphate dehydrogenase (gapdh) transcripts from isolated nuclei or whole embryos. Myl7 and vmhc are expressed exclusively in the myocardium, while slu7 is expressed ubiquitously. Error bars represent standard deviations from triplicate RT-qPCR experiments.

D Scatterplots of log2 fold differences between biotagged biological duplicates for whole embryo (bactin nuclear and bactin ribo) cardiomyocyte (myl7 nuclear) and neural crest (sox10 nuclear and sox10 ribo) samples. E Table presenting correlation coefficients to all possible pairwise comparisons of replicates/samples.
Figure S4. Neural crest identity of sox10 Biotagged nuclei. Related to Figures 5 and 6.
**A** Volcano plot of differential expression analysis of ncBirA;nuAvi samples compared to whole embryo transcriptome, demonstrating the relationship between the $p$-value and log-fold change (red=enriched, blue=decreased in ncBirA;nuAvi samples) for the 9544 genes differentially expressed. **B** Heatmap of enriched genes. **C** Heatmap of depleted genes. Brackets highlight cluster of genes with variation between sox10 biotagged samples. Heatmaps reveal differences in the sox10 nuclear transcriptome compared to the whole embryo replicates. **D** Gene set enrichment analysis for the 3767 genes enriched (red) and the 5414 genes decreased (blue) in sox10 nuclear transcriptome compared to whole embryo transcriptome. The maximum node contains 294 genes enriched for the Wnt signaling pathway, while the minimal node contains 6 genes decreased for general mRNA splicing machinery in the sox10 nuclear samples. Size of node corresponds to number of genes in each gene set. The $p$-values are presented by color saturation; the numbers and their corresponding pathway for each node are listed below. **E** Barplot representation of average FPKM expression values across replicates of sox10 nuclear polyA enriched transcriptome for all 236 neural crest genes as defined by *in situ* hybridization analysis from 0-24hpf in Zfin expression database. **F** List of 236 genes expressed in neural crest cells by 24hpf as defined by *in situ* hybridization as obtained from Zfin zebrafish gene expression database. Enriched genes (red text), decreased genes (blue text) and not differentially expressed (black text).
Figure S5: K-mean analysis identifies distinct clusters with bidirectional transcription at sites of open promoters. Related to Figure 5.
A Heatmap of $k$-mean clustering identifying 10 distinct clusters with varying levels of short bidirectional transcripts at open promoters. B-C Gene ontology (GO) terms for biological processes enriched for subclusters of genes with bidirectionally transcribed TSS in $sox10$ nuclear dataset (B) and subclusters of genes with transcription at TSS only in sense direction (C). Bidirectionally transcribed loci associate with GO terms reflecting various developmental processes including eye and sensory morphogenesis, neurogenesis and cellular differentiation with high statistical significance ($p<0.01$), while loci not exhibiting bidirectional transcription at TSS associate with various metabolic processes and have no developmental feature ($p<0.01$). D-E Donut charts comparing functional classification according to protein class (D) and biological function (E) between genes with unidirectionally (inner donut chart) and bidirectionally transcribed TSSs (outer donut charts). Most significant difference between the two gene clusters is much larger number of transcription factors amongst bidirectionally transcribed loci (D) as well as increase in number of loci associated with developmental processes (E).
Figure S6. Identification of cis-regulatory modules (CRMs) through bidirectional transcription at sites of open chromatin (ATAC-peaks). Related to Figure 6.
A Heatmap showing all k-mean clusters from linear enrichment of mapped reads from sox10 and bactin nuclear and ribosomal datasets associated across regions of open chromatin defined by ATAC-seq (±1.5kb per region). Ten clusters, totaling 65,458 distal open chromatin regions, were identified with approximately half of the ATAC-regions not associated with active transcription (cluster 4; 30,669 peaks). Three clusters show bidirectional transcription in sox10 nuclear dataset (red box, also presented in Figure 6B of main text). Two groups of 2 reciprocal clusters each (clusters 5-6 and 7-8; total 12,230 CRMs) show associated bidirectional transcription in sox10 and bactin nuclear dataset. These clusters most likely reveal ubiquitous enhancers. Two clusters contained elements with associated transcripts in both nuclear and ribosomal compartment (clusters 9-10; total 2,927 elements). B Violin plot visualising the distribution of ATAC-seq signal for NC-specific bidirectionally transcribed CRMs (Clusters 1, 2 and 3) and for non-transcribed accessible regions (Cluster 4). Although there’s a greater variation in signal level distribution for the non-transcribed cluster 4, the median value of ATAC signal on transcribed and non-transcribed regions is similar. C Annotated genes ranked by the number of associated CRMs. D Cumulative frequency plot quantifying number of associated enhancers (identified from cluster 1 and 2, total of 6332 genes) per expressed gene (based on ATAC_TSS dataset). ~47% of open annotated loci were associated with at least 2 CRMs, ~25% with 3 or more elements and 15% with 4 or more elements.
Figure S7. Comparative genome-wide profiles of nuclear and ribosomal transcripts in *sox10, myl7* subpopulation and whole embryo reveal subcellular compartment and cell-type differences. Related to Figure 7.
A Visualizations of the *bactin* and *sox10* nuclear transcriptomes on a global scale using deepTools heatmapper and profiler demonstrate a unique architecture characteristic of pervasive transcription that is enriched (with *bactin* ubiquitous control shown in replicates) evenly across untranslated and translated regions of annotated gene bodies demarcated by transcription start sites (TSS) and transcription end sites (TES). This enrichment is lost in the *sox10* ribosomal transcriptome signature (again, with *bactin* ubiquitous control shown in replicates), where higher levels of transcription in the central, translated regions of gene bodies define the transcriptional structure. As expected, the ribosomal transcriptional signature is highly similar to the profile of whole *sox10*-expressing cells that were obtained via FACS, where majority of transcripts (~90%) recovered in this manner are cytosolic. B Venn diagram showing RNA species distribution between nuclear and ribosomal dataset from whole embryo and C the corresponding RNA species found in the respective subcellular compartments with color-code referring to the different RNA species in lower panel. D Venn diagram comparing transcript differences between *myl7*, *sox10* and whole embryo nuclear datasets. E Heatmap of 51 differentially expressed lncRNAs ($p<0.05$) in the *myl7* versus *bactin* nuclear datasets (26-30hpf). 14 non-coding transcripts common between NC and myocardial differentially expressed lncRNAs are labelled in red (depleted) and blue (enriched) in the *myl7* versus *bactin*-expressing nuclei. F Venn diagram comparing NC and *myl7* differentially expressed lncRNAs.
Table S1. Biotagging driver and effector lines. Related to Figures 1 and 2.

| Biotagging BirA driver          | Allele | Abbreviation | Expression | Note                                                                 |
|--------------------------------|--------|--------------|------------|----------------------------------------------------------------------|
| Tg(bactin:BirA-membCherry)     | ct709a | ubBirA       | ubiquitous |                                                                     |
| Tg(bactin:NLS-BirA-membCherry) | ct710a | ubBirA(NLS)  | ubiquitous |                                                                     |
| Tg(ubiq:NLS-BirA-Cherry)       | ox114  | ubBirA(ubiq-NLS) | ubiquitous |                                                                     |
| Tg(ubiq:BirA-Cherry)           | ox115  | ubBirA(ubiq) | ubiquitous |                                                                     |
| Tg(sox10:BirA-membCherry)      | ct706a | ncBirA       | cranial neural crest derivatives | female-exhibit ubiquitous maternal expression |
| Tg(sox10:BirA-membCherry)      | ct706b | ncBirA(b)    | cranial and trunk neural crest derivatives |                                                                     |
| Tg(zic2a:NLS-BirA-membCherry)  | ct708a | hbBirA       | presumptive telencephalon and diencephalon, hindbrain primary neurons | expression gone by 2dpf |
| Tg(my7:BirA-membCherry)        | ct704a | myoBirA      | myocardium |                                                                     |
| Tg(my7:NLS-BirA-membCherry)    | ct705a | myoBirA(NLS) | myocardium |                                                                     |
| Tg(kdrl:BirA-membCherry)       | ct703a | endoBirA     | vascular endothelium |                                                                     |
| TgBAC(sox10:BirA-Cherry)       | ox104a | ncBirA(BAC)  | neural crest and derivatives | female may exhibit ubiquitous maternal expression |

| Biotagging Avi effector        | Allele | Expression | Note |
|--------------------------------|--------|------------|------|
| Tg(bactin:Avi-Cerulean-RanGap)| ct700a | nucAvi(bact)| ubiquitous |
| Tg(bactin:RanGap-Cerulean-Avi) | ct701a | nucAvi(bact-Cterm) | ubiquitous |
| Tg(ubiq:Avi-Cerulean-RanGap)  | ox113a/b | nucAvi(ubiq) | ubiquitous |
| Tg(bactin:Avi-Cerulean-Rpl10) | ox111  | riboAvi(bact)| ubiquitous |
| Tg(ubiq:Avi-Cerulean-Rpl10)   | ox112  | riboAvi(ubiq) | ubiquitous |
| Gene Symbol | Accession Number | Description |
|-------------|-----------------|-------------|
| ENSDARG00000035957 | gmnn | unigene:786571 gsnRNA isoform X10.1 |
| ENSDARG00000010655 | ppm1k | unigene:1644845 gsnRNA isoform X10.1 |
| ENSDARG00000033733 | ntn1a | unigene:1640066 gsnRNA isoform X10.1 |
| ENSDARG00000078366 | robo2 | unigene:1640706 gsnRNA isoform X10.1 |
| ENSDARG00000068567 | shha | unigene:1637577 gsnRNA isoform X10.1 |
| ENSDARG00000007369 | tcf7l1b | unigene:820396 gsnRNA isoform X10.1 |
| ENSDARG00000076393 | tmem65 | unigene:1639590 gsnRNA isoform X10.1 |
| ENSDARG00000006272 | mpp5a | unigene:1639166 gsnRNA isoform X10.1 |
| ENSDARG00000061052 | nipblb | unigene:1638942 gsnRNA isoform X10.1 |
| ENSDARG00000062909 | furina | unigene:1639137 gsnRNA isoform X10.1 |
| ENSDARG00000042535 | actc1a | unigene:1638111 gsnRNA isoform X10.1 |
| ENSDARG00000003938 | rpa1 | unigene:1640052 gsnRNA isoform X10.1 |
| ENSDARG00000011027 | fgfr1a | unigene:1640803 gsnRNA isoform X10.1 |
| ENSDARG00000056964 | ilk | unigene:1639359 gsnRNA isoform X10.1 |
| ENSDARG00000070670 | crip2 | unigene:1638566 gsnRNA isoform X10.1 |
| ENSDARG00000058449 | arnt2 | unigene:1639111 gsnRNA isoform X10.1 |
| ENSDARG00000017803 | gsk3b | unigene:1639877 gsnRNA isoform X10.1 |
| ENSDARG00000030213 | sap30l | unigene:1638476 gsnRNA isoform X10.1 |
| ENSDARG00000016744 | fbn2b | unigene:1639193 gsnRNA isoform X10.1 |
| ENSDARG00000071395 | camk2g1 | unigene:1639473 gsnRNA isoform X10.1 |
| ENSDARG00000058115 | fgfr2 | unigene:1638536 gsnRNA isoform X10.1 |
| ENSDARG00000040334 | mat2aa | unigene:1638835 gsnRNA isoform X10.1 |
| ENSDARG00000077777 | tmsb4x | unigene:1638735 gsnRNA isoform X10.1 |
| ENSDARG00000002710 | ncl | unigene:1638291 gsnRNA isoform X10.1 |
| ENSDARG00000021124 | cfl1 | unigene:1638453 gsnRNA isoform X10.1 |
| ENSDARG00000077004 | aldh1l1 | unigene:1639516 gsnRNA isoform X10.1 |
| ENSDARG00000000529 | ofd1 | unigene:1638600 gsnRNA isoform X10.1 |
| ENSDARG00000035427 | surf6 | unigene:1639229 gsnRNA isoform X10.1 |
| ENSDARG00000075621 | birc5a | unigene:1638860 gsnRNA isoform X10.1 |
| ENSDARG00000071465 | lrrc39 | unigene:1639268 gsnRNA isoform X10.1 |
| ENSDARG00000028067 | bnip3lb | unigene:1638978 gsnRNA isoform X10.1 |
| ENSDARG00000033999 | cyp26a1 | unigene:1639063 gsnRNA isoform X10.1 |
| ENSDARG00000069090 | atp6v0d1 | unigene:1639135 gsnRNA isoform X10.1 |
| ENSDARG00000073765 | gna13a | unigene:1639267 gsnRNA isoform X10.1 |
| ENSDARG00000026348 | csad | unigene:1638459 gsnRNA isoform X10.1 |
| ENSDARG00000036548 | s1pr2 | unigene:1638501 gsnRNA isoform X10.1 |
| ENSDARG00000041421 | taz | unigene:1638715 gsnRNA isoform X10.1 |
| ENSDARG00000012397 | eya4 | unigene:1638426 gsnRNA isoform X10.1 |
| ENSDARG00000030722 | xirp1 | unigene:1638589 gsnRNA isoform X10.1 |
| ENSDARG00000026784 | robo1 | unigene:1638535 gsnRNA isoform X10.1 |
| ENSDARG00000045097 | bmpr1ab | unigene:1639165 gsnRNA isoform X10.1 |
| ENSDARG00000003398 | rbpja | unigene:1638930 gsnRNA isoform X10.1 |
| ENSDARG00000037924 | gna13b | unigene:1638297 gsnRNA isoform X10.1 |
| ENSDARG00000020201 | cdc73 | unigene:1638416 gsnRNA isoform X10.1 |
| ENSDARG00000029370 | ankrd6b | unigene:1638428 gsnRNA isoform X10.1 |
| ENSDARG00000034624 | nuf2 | unigene:1638665 gsnRNA isoform X10.1 |
| ENSDARG0000006603 | csrp1a | unigene:1638772 gsnRNA isoform X10.1 |
| ENSDARG00000020334 | ptpn11a | unigene:1638410 gsnRNA isoform X10.1 |
| ENSDARG00000009418 | mef2cb | unigene:1638418 gsnRNA isoform X10.1 |
| ENSDARG00000021232 | nkx2.7 | unigene:1638399 gsnRNA isoform X10.1 |
| ENSDARG00000024894 | tbx5a | unigene:1638397 gsnRNA isoform X10.1 |
| ENSDARG00000008398 | cacna1c | unigene:1638394 gsnRNA isoform X10.1 |
| ENSDARG00000004405 | snx10a | unigene:1638396 gsnRNA isoform X10.1 |

Table S2. Genes overrepresented in myl7 vs bactin nuclei. Related to Figure 4.
Supplemental Experimental Procedures

Zebrafish maintenance and strains.

This study was carried out in accordance to procedures authorized by the UK Home Office in accordance with UK law (i.e. Animals (Scientific Procedures) Act 1986) and the recommendations in the Guide for the Care and Use of Laboratory Animals (US). Adult fish were maintained as described (Westerfield, 2000). Wild-type embryos for transgenesis were obtained from AB or AB/TL mix strains.

BirA cassette design.

The tripartite biotinylation driver expression cassette consists of open reading frame (ORF) of bacterial biotin ligase, BirA, preceded by 3xHA sequence for protein detection and separated from the membrane-tethered mCherry fluorescent reporter (membCherry) by a short sequence encoding the ribosome-skipping peptide of *Thosea asigna* virus (2A, Fig.1B-D). The membrane localization signal for membCherry was derived from the last 20 amino acids of human Ras (Apolloni et al., 2000). The biotinylation driver construct was generated by fusion PCR with 3XHA-BirA and membCherry templates, using intervening overlap sequence between the two to encode 2A sequence. The 20 bp overlap was built into the primers used for amplification of the templates. We have generated BirA drivers with both cytoplasmic and nuclear cellular localization (NLS) to enable biotinylation of both cellular component-associated Avi-tagged proteins (nuclear envelope and cell membrane) and Avi-tagged intra-nuclear factors, respectively. The full list of generated biotinylation drivers is shown in Table S1. Full sequences of the plasmids are available through NCBI and Addgene (https://www.addgene.org/Tatjana_Sauka-Spengler/).

Generation of Biotagging transgenic drivers.

Tol2-mediated transgenesis for driver lines: Biotagging transgenic drivers were created using conventional Tol2-mediated zebrafish transgenesis (Kawakami, 2004). BirA expression cassette was placed under the control of previously published proximal enhancers and tissue- or cell-specific promoters (Table S1) to create defined expression patterns with the entire expression module flanked by Tol2 transposable elements. The presence of the 2A peptide allowed for simultaneous expression of BirA and the fluorescent membrane-Cherry (membCherry) reporter for screening and imaging purposes (Fig.1B-D, B’-D’). The transgenes were generated by co-injecting 80 pg DNA expression constructs and 40 pg of *tol2* mRNA into single cell embryos. The injected embryos (mosaic F0 generation) were raised to reproductive age, out-crossed to wildtype adults and the F1 offspring screened for proper expression of fluorescent reporter. F1 carriers were raised for future experiments. Subsequent generations of transgenes are maintained as out-crosses to wildtype adults to ensure single copy transgenic propagation.

Tol2-mediated Biotagging transgenic drivers for expression of BirA produced in this study include four tissue-specific lines: Tg(sox10:BirA-2A-membCherry)ct706a (ncBirA) expressing BirA in delaminating and migrating neural crest under control of the *sox10* promoter (Carney et al., 2006) (Fig.1B,B’; Fig.2G); Tg(zic2a:BirA-2A-membCherry)ct708a (hbBirA) in the neural plate border cells under the control of the *zic2a/zic5* enhancer (Nyhholm et al., 2007) (Fig.1D,D’); Tg(myl7:BirA-2A-membCherry)ct704a (myoBirA) in the myocardium of the developing heart (Huang et al., 2003) (Fig.1C,C’) and Tg(kdrl:BirA-2A-membCherry)ct703a (endoBirA) in endothelium of the circulatory system (In et al., 2005) (not shown). In addition, ubiquitous BirA driver lines included Tg(bactin:BirA-2A-membCherry)ct709a (ubBirA) and Tg(ubiq:BirA-2A-membCherry)ox115 (ubBirA(ubiq)) (Higashijima et al., 1997; Mosimann et al., 2011) (Fig.1E). Test experiments, using homozygote Tg(bactin:BirA-2A-membCherry)ct709a (ubBirA) embryos (Fig.1E), which express BirA at very high levels, show no developmental defects and can reproduce, indicating that expression of BirA is not toxic in zebrafish.

BAC recombineering for driver lines: The availability of well characterized *cis*-regulatory modules (CRMs) for BirA drivers can limit the application of the Tol2-mediated transgenes driver approach. Recombineered BACs, containing gene-associated regulatory elements, can serve as an alternative to transgenic BirA drivers using known CRMs. To generate Biotagging BAC drivers we replaced the first coding exon of the gene of interest with a BirA cassette (Fig.1G) and used Tol2-mediated transgenesis to integrate recombineered BACs into zebrafish genome.
To achieve this, we generated the donor cassette containing HA-tagged BirA ORF, separated from mCherry reporter by ribosomal-skipping peptide (2A) and terminating with polyA, followed by FRT site-flanked Kanamycin selection gene and recombined it into the selected BAC backbone using lambda prophage homologous recombination system available in the SW105 bacterial background (http://ncifrederick.cancer.gov/research/brb/productDataSheets/recombineering/bacterialStrains.aspx), according to the previously published protocol (Yu et al., 2000). To increase the efficiency of transgenesis and enable single-copy integration into zebrafish genome, the BACs were also modified to include the long terminal repeats (LTRs) of the Tol2 transposon. The iTol2-Amp cassette, containing the Ampicillin expression construct flanked by inverted Tol2 recombination arms was amplified according to published protocols (Abe et al., 2011), using iTol2-Amp plasmid and previously described primers:

\[
\text{pIndigobacitol2}_\text{fw:}
\text{TTCCTGTGTATTGTCGGAGAACAAATGGAAGTCCGAGCTCATCGCTCCCTGCTCGAGCCGGCCCAAGTG}
\]
\[
\text{pIndigobacitol2}_\text{rev:}
\text{CCCGCCAACACCCGCTGACGCGAACCCCTTGCGGCCGCATATTATGATCCTCTAGATCATGCATTCTGATCAGATCT}
\]

and recombined into the sox10 locus-containing BAC DKEY-201F15, with pIndigo backbone. We have also created an extended iTol2-Amp cassette for integration of Tol2 arms into BAC clones with the pTARBAC backbone. The new iTol2-Amp cassette containing long homologous recombination arms (5'arm-224 bp and 3'arm-221 bp) that flank loxP sites on the pTARBAC backbone can be amplified using pTARBAC\_loxP\_5':

GCTGTCGGAATGGACGATA and pTARBAC\_loxP\_3':

GCAAGTATTGACATGTCGTCGT primers and recombined using procedures described above. 100-200 pg of recombinant BAC DNA was co-injected with 50-100 pg of tol2 mRNA into one cell-stage embryos to generate F0 generation. Potential F0 founders were raised, outcrossed and the F1 clutches screened for mCherry expression. Selected positive F1 embryos were raised for future experiments.

Using a BAC containing the sox10 locus we successfully overcame variation in expression patterns obtained from the conventional sox10 transgenics (Fig.S1A-C). We detected strong neural crest expression of mCherry in the F1 offspring from 3 out of 9 screened TgBAC(sox10:BirA-Cherry)\(^{\text{ox104a}}\) (ncBirA(BAC)) F0 adults (Fig.1G-H, Fig.S1A). All the F1 embryos from the three independent founders showed consistent neural crest-specific mCherry expression (Fig.1G-H, Fig.2I). This is in contrast to sox10 BirA (ncBirA) drivers, obtained using conventional proximal promoter transgenesis, which often exhibited variability in expression patterns between different founders due to integration position effects (Fig.S1B,C). However, we exploited this variability to generate a number of biotagging drivers that enable profiling of specific subpopulations of sox10-expressing cells (Fig.S1B,C).

**Differences between Biotagging with Transgenic and with BAC drivers:** Biotagging transgenic drivers generated via routine transgenesis approaches in zebrafish express BirA-membCherry cassette under the control of the minimal promoter and proximal enhancers (Fig.1B-D). Using this approach, it is typical to observe a large variability in expression patterns in different founders produced with the same expression construct (Fig.S1B,C), suggesting that such expression cassettes are very sensitive to position effects, and that their activity is strongly influenced by their genomic integration sites. Exploiting this variability, we generated a number of Biotagging drivers that enable profiling of specific subpopulations of sox10-expressing cells (Fig.S1B,C). However, many of the sox10 drivers do not recapitulate the full sox10 gene expression pattern characterized using in situ hybridization.

Biotagging BAC drivers are much more consistent with all founders always showing reproducible expression patterns (Fig.S1A). Moreover, they are virtually insensitive to position effects and identified genomic integrations were never found to be silent. In majority of cases, the expression of BirA transgene from the BAC backbone results in robust and strong endogenous-like level of expression.
Generation of Avi effector lines.

To generate zebrafish transgenic effector lines, we used the ubiquitous zebrafish beta-actin2 (βactin) and ubiquitin (ubiq) promoters to drive expression of Avi-tagged fusions that associate with different cellular compartments: (i) Avi-tagged protein containing Cerulean protein fused to the carboxy-terminal domain of avian Ran GTPase-activating protein 1 (RanGap1), targeted to the outer nuclear envelope, for use in INTACT procedure and (ii) Avi-tagged Rpl10a to biotinylate the polyribosomes, for use in TRAP approach.

Generating Avi-RanGap lines: To generate Avi-RanGap (nucAvi) lines, in which the nuclear envelope is specifically Avi-tagged we used C-terminal domain of the avian RanGap protein because fusions with the equivalent region from the zebrafish RanGap resulted in recombinant protein that inconsistently associated with nuclear envelope, displaying much broader cellular distribution, and affecting normal development (construct resulted in embryonic lethality when injected at high concentrations of 100 pg of mRNA per embryo, data not shown). Protein domain analysis of chicken and zebrafish RanGap domains using SMART tools (http://smart.embl-heidelberg.de) indicated that the chicken but not the zebrafish RanGap C-terminal region contains a Ran Binding Domain, which associates with nuclear pore complexes [Mahajan et al., 1997] [Rose and Meyer, 2001], to directly bind Avi-Cerulean-RanGap fusion and localize it to the outer nuclear envelope (Fig.3A, B). We generated transgenic lines expressing the both N and C-terminal Avi fusion of avian RanGap (Tg(bactin:Avi-Cerulean-RanGap)ct700a (nucAvi(bact)) and Tg(bactin:RanGap-Cerulean-Avi)ct701a (nucAvi(bact)-Cterm)). They show similar localization to the outer nuclear envelope (Fig.S1F-I).

All N-terminal effector constructs employed a modified Avi tag (14 aa), followed by a 7 aa-peptide specifically recognized and cleaved by Tobacco Etch Virus protease, generated by Strouboulis lab [Driegen et al., 2005] (Fig.S1D, TeV in green). Inclusion of a protease cleavage sequence adjacent to biotin acceptor peptide helps reduce the non-specific background resulting from streptavidin bead pull-down of endogenously biotinylated proteins. This is particularly useful in analysis of protein complexes by Mass Spectrometry, where the biotin-tagged target protein and its interacting partners can be specifically released from the streptavidin beads by TeV cleavage. C-terminal Avi-tag is preceded by TeV sequence (Fig.S1E).

To assess the possible steric effects of the Avi-tag on the localization of the RanGap fusion protein and its availability for biotinylation, we have created two versions of Avi-tagged RanGap effector lines, Avi-Cerulean-RanGap and RanGap-Cerulean-Avi (Fig.S1D,E). Both versions localized to the outer nuclear envelope (Fig.S1F-I) and are interchangeable; however, we preferentially use the Tg(bactin:Avi-Cerulean-RanGap)ct700a (nucAvi(bact)) line, where the RanGAP domain in the protein fusion is located at the C-terminus, similar to the full-length RanGap protein. To select for the most ubiquitously, even expressing Avi-tagged RanGap effectors, the offspring of multiple founders was screened by confocal microscopy.

Generating Avi-Rpl10 lines: To enable isolation of the polyribosomes, we established lines that label a component of polyribosomes, Rpl10a [Tryon et al., 2013], with the Avi-tag Tg(bactin:Avi-Cerulean-Rpl10)ox111 (riboAvi(bact)) and Tg(ubiq:Avi-Cerulean-Rpl10)ox112 (riboAvi(ubiq)). While lines established with the ubiquitin promoter resulted in higher expression of Avi-Cerulean-Rpl10, we also found that high expression of Rpl10 resulted in embryonic lethality (data not shown). Only founders that resulted in Avi-Rpl10 lines showing no developmental defects that could reproduce were maintained and used for subsequent profiling.

Preparation of Streptavidin beads for RNA procedures.

As Streptavidin Dynabeads (Invitrogen, cat. no.11205D) are not supplied in RNase-free solutions, 250 μg of M-280 or MyOne T1 Streptavidin-coated beads was transferred to a microcentrifuge tube, separated from supernatant using magnetic stand (DynaMag™-2 magnet from Invitrogen, cat. no.12321D) and washed twice with 1 mL of Solution A (DEPC-treated 0.1 M NaOH, DEPC-treated 0.05 M NaCl), with 2-3 minutes on a nutator, followed by at least 3 minutes on the magnetic stand. The beads were subsequently washed once in Solution B (DEPC-treated 0.01 M NaCl), re-suspended in NPB and moved to a new RNase-free low-binding tube, until the nuclei suspension was ready. Prior to incubation with cell nuclei/polysomes, the beads were captured using the magnetic stand, the supernatant removed, and replaced with the nuclei/polysomes suspension.
Nuclei Isolation.

We optimized the nuclei purification protocol for the highest yield by testing a number of lysis buffers and found the following to give the highest consistent yield of nuclei per embryo. Zebrafish embryos (~100-350 embryos per pulldown experiment) expressing both biotinylation driver and Avi-RanGap effector alleles (nucAvi) in specific cell types were dechorionated and washed in hypotonic Buffer H (20 mM HEPES (pH 7.4), 1.5 mM MgCl₂, 10 mM KCl, 1 mM DTT and 1X cOmplete™ protease inhibitor), supplemented with 0.01% Tricaine. Embryos then were re-suspended in 1 mL/50 embryos Buffer H and transferred to a Dounce homogenizer (2 or 7 mL Kontes Glass Co, Vineland, NJ). Embryos were dissociated with a sequence of 20 strokes using loose fitting pestle A, incubated on ice for 15 minutes, followed by 60 strokes of tight fitting pestle B (3 x 20 strokes, pausing 5 minutes on ice after each set of 20 strokes) to allow for lysis of cell membranes. Cells were checked for lysis by visualizing cells with 1:1 dilution of Trypan blue on microscope. Nuclei were collected by centrifugation for 10 minutes at 2,000g and re-suspended in 1 mL of nuclei purification buffer (NBP: 10 mM HEPES (pH 7.4), 40 mM NaCl, 90 mM KCl, 0.5 mM EDTA, 0.5 mM spermidine, 0.15 mM spermine, 1 mM DTT and 1X cOmplete™ protease inhibitor (Roche, cat. no.05892791001)). To purify nuclei, the suspension was incubated with 250 µg (1 x 10⁷ beads) of M-280 Streptavidin-coated Dynabeads prepared for RNA procedures (see preparation of beads) with rotation at 4 °C for 30 minutes. The nucleus-bead suspension was further diluted with NPB containing 0.1% Triton X-100 (NPBt) to a final volume of 20 mL. A flow-based setup was devised as previously published (Deal and Henikoff, 2011) using 10 mL plastic serological pipette (VWR, cat. no.89130-898 or BD Falcon, cat. no.3575551) attached to a 1 mL micropipette tip (Gentaur Reach Pipet Tip, cat. no.24-1635R or Rainin, cat. no.RT-L1000S), both pre-treated with NPB + 1% BSA for 10 minutes to avoid material loss. To capture the specifically biotinylated nuclei, the 1 mL pre-coated pipette tip was attached to the MiniMACS separator magnet (OctoMACS Separator, Miltenyi Biotec, cat. no.130-042-109) and the diluted suspension of nuclei-beads was allowed to flow through the setup. The flow rate was set to ~0.75 mL min⁻¹ using a two-way stopcock (Bio-Rad, cat. no.732-8102) or a T-valve (from Bio-Rad Low-pressure fittings kit, cat. no.731-8220), connected at the end of the tip via a short piece of Tygon tubing (Fisher Scientific, cat. no.14-169-1C). For flow rates faster than 0.75 mL min⁻¹, the flow-through was drawn up with the same serological pipette, without removing the pipet tip from the magnetic separator and the suspension was allowed to drip through the tip one more time to maximize recovery efficiency. Subsequently, pipet tip was removed from the stand, attached to p1000 pipetman and beads and nuclei were released from the wall of the tip by repeated drawing of fresh 20 mL NPBt into and out of the tip and the magnetic purification was repeated as described above. At the end, beads and nuclei were eluted in 1 mL of NPBt, placed into the 1.7 mL microfuge tube and onto the magnetic stand (DynaMag™-2 magnet from Invitrogen, cat. no.12321D) to remove the supernatant. Purified nuclei-beads were re-suspended in 20 µL NPB, stained with DAPI and imaged or frozen for future use. Nuclei yield after purification were determined by staining 1:20 of the total nucleus prep with DAPI and subsequent counting of the number of bead-bound nuclei. When used for RNA extraction, the 250 µg bead-nuclei pellets were immediately dissolved in 100 µL of RNA lysis buffer, incubated at room temperature for 10 minutes and replaced onto the magnetic stand. The RNA lysis buffer containing total nuclear RNA is then removed into a fresh tube and snap-frozen for future use or for immediate extraction (see RNA extraction and library preparation).

Polysomal Isolation.

We adapted the polysomal isolation protocol from TRAP, Translating Ribosome Affinity Purification (Heiman et al., 2014) with modifications optimised for the Biotagging system. Zebrafish embryos (350 embryos) expressing both biotinylation driver and Avi-Rpl10a effector alleles (riboAvi) in specific cell types were dechorionated and washed in Cell Lysis Buffer (20 mM HEPES (pH 7.4), 150 mM KCl, 10 mM MgCl₂, 0.5 mM DTT, RNasin, RNaseOUT, SUPERaseIN and cOmplete™-EDTA-free protease inhibitor). Embryos were lysed in 7 mL of Cell Lysis Buffer without cycloheximide in a Dounce homogeniser as described in the Nuclei Isolation protocol above. The key difference is that cycloheximide was added to the cell suspension after pestle A to a final concentration of 100 µg mL⁻¹ and incubated at RT for 15 minutes before proceeding with pestle B strokes. Homogenised embryos were cleared by centrifugation at 2000g for 10 minutes at 4 °C and the post-nuclear supernatant removed into clean RNase-free 1.7 mL microfuge tubes (1 mL supernatant per tube). IGEPAL CA-630 and 07:0 DHPC were added to the supernatant to a final concentration of 1% each and mixed by inverting tubes gently 10 times. Following incubation on ice for 5 minutes, the supernatant was cleared by centrifugation at 20,000g for 10 minutes at 4 °C.
To purify polysomes, the post-mitochondrial supernatant was removed and added to 250 μg (2.5 x 10^8 beads) of MyOne T1 Streptavidin-coated Dynabeads (1 mL supernatant per 250 μg beads) prepared for RNA procedures (see preparation of beads) with rotation at 4 °C for 1 hour. The tubes containing polysomes-beads suspension were placed onto a magnetic stand (DynaMag™-2 magnet from Invitrogen, cat. no.12321D) to remove the unbound lysate. The pelleted polysomes-beads were washed four times (changing tubes in between washes to minimise background) in the cold room with High Salt Buffer (20 mM HEPES (pH 7.4), 350 mM KCl, 10 mM MgCl₂, 0.5 mM DTT, rRNasin, RNaseOUT, SUPERaseIN, 100 μg ml⁻¹ cycloheximide and 1% IGEPA/CA-630) by pooling 500 μg beads per 1 mL High Salt Buffer. After the final wash, the tubes with polysomes-beads were placed onto a magnetic stand to remove the High Salt Buffer. When used for RNA extraction, the 500 μg polysomes-beads pellets were immediately dissolved in 200 μL of RNA lysis buffer, incubated at RT for 10 minutes and replaced onto the magnetic stand. The RNA lysis buffer containing polysomal-bound RNA is then removed into a fresh tube and snap-frozen for future use or for immediate extraction (see RNA extraction and library preparation).

**FACS and ATAC.**

The sox10-expressing cells were isolated from TgBAC(sox10:BirA-mCherry)^x104a (ncBirA(BAC)) embryos at 16ss using Fluorescence Activated Cell Sorting (FACS). Prior to FACS embryos were dissociated using 20 μg ml⁻¹ collagenase in 0.05% Trypsin/0.53 mM EDTA/1xHBSS buffer to obtain single cell suspensions. Reaction was stopped in 10 mM HEPES/0.25% BSA/1xHBSS buffer and mCherry-positive neural crest cells were sorted using BD FACSAria Fusion System. Sorted cells were spun down and washed in PBS and immediately used in ATAC procedure. Tagmentation was performed as previously described [Buenrostro et al., 2015]. Fragment size was verified using Tapestation (Agilent) and libraries were quantified using KAPA Library Quant Kit for Illumina Sequencing Platforms (KAPABiosystems).

**Western blot analysis.**

Protein extract was obtained from zebrafish embryos at specified time-points. Embryos were de-yolked, lysed with a Dounce homogenizer (Pestle A) in hypotonic buffer (20 mM HEPES (pH 7.9), 1.5 mM MgCl₂, 10 mM KCl, and 1 mM DTT) with protease inhibitors, and centrifuged at maximum speed to obtain cytoplasmic fraction. Nuclear fraction was obtained by lysis of remaining nuclei pellet in nuclear lysis buffer (20 mM HEPES (pH 7.9), 1.5 mM MgCl₂, 0.2 mM EDTA, 20% glycerol, 420 mM KCl, 0.4 mM PMSF) with protease inhibitors using Pestle B followed by centrifugation at maximum speed. Detection was performed with anti-HA antibody (Roche (cat. no.12CA5), 1:1000), rabbit anti-GFP antibody (1:1000, Torrey Pines Biolabs, Houston/TX, www. chemokine.com, used for detection of Cerulean) and Streptavidin-HRP Conjugate (used for detection of biotin). GFP and HA-tag were detected using standard Western blot procedure, while biotinylated proteins were detected using a modified procedure. After the transfer, the blots were blocked for 1 hr in 5% BSA/1X TBST (20 mM Tris, 137 mM NaCl, 0.2% TWEEN-20) and incubated for 1 hr at room temperature with Streptavidin–HRP Conjugate (NEL750, Perkin Elmer,1:10,000). Filters were then washed 6 times for 20 minutes in 1X TBS (20 mM Tris, 137 mM NaCl) + 0.3% Triton X-100 and signal was detected using ECL Plus Western Blot Detection Reagent (GE Healthcare Life Sciences, cat. no. RPN2132).

**Quantitative Real-Time PCR analysis.**

Two-step qRT-PCR was performed using ABI’s Sybr-Green RT-PCR system (Applied Biosystems). Briefly, RNA was extracted from nuclei isolated from 48hpf embryos double transgenic for the Biotagging myocardial driver allele, Tg(myl7:BirA-2a-membCherry)^x704a (myoBirA) and the Avi effector allele, Tg(bactin:Avi-Cerulean-RanGap)^x700a (nuAvi(bact)) using RNAqueous Micro kit (see RNA extraction and library preparation). cDNA was synthesized with reverse transcriptase (SuperScript II RT, Invitrogen) using random hexamers for priming. Reverse transcription reactions were diluted in series (1-10,000-fold) and 1 μL was amplified in triplicates on a 7000 Sequence Detection System (Applied Biosystems). Quantification was performed using the delta-delta Ct (ΔΔCt) method [Livak and Schmittgen, 2001]. Primers used for qRT-PCR are as follows: m yell (myl7_fw: AGGGGGAAACTGCTCAAAG and myl7_rev: TGATACTCCATCCCGTTC), vmhc (vmhc_fw: TCGTCAGTCGTAAGAGTGAC and vmhc_rev: GGCTCATAGAAGGAAGGTGAA), slu7
Quality control for Nuclei and Polysomal Isolation, RNA extraction, and library preparation.

Maximal efficiency of nuclei and polysomal isolation is highly dependent on complete lysis of cells with Buffer H (nuclei isolation) and Cell Lysis Buffer (polysomal isolation). Hence, it is essential that embryos are lysed at a ratio of 50 embryos (26hpf or younger) per 1 mL of buffer with 20 strokes of pestle A and 60 strokes of pestle B. Care must be undertaken that pestle B is not faulty and has the appropriate small clearance for efficient cell lysis. Quality of cell lysis for each experiment was determined by RNA extraction of the unbound fraction. For nuclei isolation, the flow-through containing unbound nuclei (the ‘lysate’) was pelleted by centrifugation at 2000g for 10 minutes at 4 °C and dissolved in 400 µL RNA lysis buffer. The sample was placed onto a magnetic stand to remove residual Dynabeads; RNA lysis buffer containing total RNA from the lysate was processed in the same manner. Similarly, for polysomal isolation, the supernatants containing unbound polysomes were cleared by centrifugation at 7000g for 10 minutes at 4 °C, the supernatant removed until about 200 µL is left, followed by addition of 800 µL RNA lysis buffer.

Both purified and unbound nuclei and polysomes were lysed and RNA pools extracted using RNAqueous Micro Scale Total RNA Isolation Kit (Ambion cat. no.AM1931), genomic DNA was removed by 20 minutes of rDNasel (provided with Ambion cat.no.AM1931) treatment. Before library production, quality of the RNA was assayed using Agilent RNA 6000 Pico kit (Agilent Technologies, cat. no.5067-1513) on the Agilent 2100 Bioanalyzer, as specified by manufacturer. We ensured that in both nuclei and polysomal isolation experiments, the Bioanalyzer profiles for experimental and lysate samples were highly similar and displayed the expected quantity ratios (i.e. much higher amount of RNA in the lysate sample compared to experimental sample). Non-directional sequencing libraries after polyA-selection of RNA transcripts (NEBNext® Poly(A) mRNA Magnetic Isolation Module, NEB) were built using NEBNext Ultra RNA library kit for Illumina (NEB). For directional RNA-sequencing, 30-50 ng of total nuclear RNA and 40-50 ng of total polysomal RNA were first enriched by ribodepletion using Ribo-Zero™ Magnetic Kit (EPICENTRE). Subsequently sequencing libraries were prepared using Stranded RNA-Seq Library Preparation Kit (KAPABiosystems), according to manufacturer’s instructions. Deep sequencing was performed on HiSeq2500 or Nextseq500 illumina platforms. Biological duplicates were generated for each experimental condition and pairwise comparison performed on biological duplicates to ensure high quality of sequence data for analysis (Fig.S3D). cDNA libraries for RT-PCR validation were generated using Superscript II RT and random hexamer priming (Life Technology, cat. no.18064-014). To directly compare different total RNA isolation protocols - biotagged nuclei, biotagged ribosomes and FACS total RNA yield was quantified per batches of 100 embryos. Number of positive cells/organelles per embryos recovered was deduced and calculated using previously defined standards of 1 pg RNA/cell for Avi-RanGap, 0.05 pg RNA/cell for Avi-Rpl10 and cell counts from FACS experiment.

Bioinformatics Processing.

**ATAC-Seq:** ATAC-seq data was sequenced using paired-end 40 bp run on the NextSeq500 platform. Reads were trimmed for quality using sickle (v 1.33) [Lohi and Fass 2011] and mapped using bowtie (v.1.0.0). Bigwig files were generated using an enhanced Perl script courtesy of Jim Hughes. Only paired reads with insert sizes larger than 100 bp were selected and reads were displaced by +4 bp and -5 bp as described previously [Buenrostro et al. 2013] and extended to a read length of 100bp. Peak calling was performed using MACS2 with --nomodel and --locl 1000 parameters [Zhang et al. 2008]. Zebrafish Ensembl gene models were extended by 100 bp in 5' of the TSS to account for gene mis-annotation. ATAC-seq peaks overlapping with extended TSSs were used to define open promoter set (ATAC_TSS). Putative cis-regulatory element set (ATAC_enhancer), was identified as ATAC peaks not overlapping with Ensembl-annotated promoter regions or exons.

**RNA-seq analysis:** RNA-Seq data was sequenced using 50 bp paired-end reads on HiSeq2000 and HiSeq2500 platforms. Whole embryo polyadenylated transcriptome at 24hpf generated by Armant and colleagues [Armant 2008].

*slu7_fw: AGAAAAAGGAGCATGCGAAAA and slu7_rev: atgcctgtgccagaaaactt) and gapdh (gapdh_fw: GATACACGGAGCACCAGGTT and gapdh_rev: CGTTGAGAGCAATACCAGCA).*
et al., 2013) was downloaded from SRA (Accession SRP014596). Reads were mapped to the zebrafish genome (Jul.2010 Zv9/danRer7 assembly) with STAR (v.2.4.2a) using default parameters (Dobin et al., 2013). Sets of BAM files incorporating reads belonging to either DNA strand were generated using custom scripts available at https://github.com/tsslab/biotagging/. Count tables were produced for Ensembl gene models using subread featureCount v1.4.5 (Liao et al., 2014) or htsseq-count for strand-specific quantification. Differential expression analysis for different gene models (ENSEMBL gene models, custom gene models for intron quantification and published lncRNA models) was performed using DESeq2 (Anders et al., 2012; Love et al., 2014). Enriched genes were selected at a p-value of 0.05 after a Benjamini-Hochberg adjustment for multiple testing. Gene set enrichment analysis was performed using the Piano package (Varemo et al., 2013) and the Panther pathway classification downloaded for zv9 version of the genome (ftp://ftp.pantherdb.org/). Transcript levels were quantified in RPKM and FPKM, as previously described (Mortazavi et al., 2008). Genes expressed at FPKM > 1 were deemed expressed. Data generated in this study submitted to GEO (GSE89670) and are also available via Daniocode consortium (http://danio-code.zfin.org/daniocode/).

**Genome-wide analysis of polyA-enriched neural crest nuclear transcripts validates the Biotagging approach:** Previous studies using the INTACT system employed polyA-based enrichment of RNA, thus harvesting the spliced portion of the nuclear transcriptome. To cross-validate our approach, we applied similar analyses to the nuclear RNA pool isolated from 24hpf neural crest cells, biotagged by crossing the Tg(sox10: BirA-2A-membCherry)^ct706a line (ncBirA) with the Avi-RanGAP effector line (nucAvi(bact)). RNA-seq libraries were prepared from polyA-selected nuclear transcripts, sequenced and analyzed. Differential expression analysis comparing polyA-selected nuclear neural crest to the whole embryo transcriptomes at 24hpf (Armant et al., 2013), identified 6580 differentially expressed genes (p<0.05), with 2918 genes significantly enriched and 3662 decreased in the sox10 nuclear samples (Fig.S4A). Biological replicates of sox10 biotagged nuclei samples were strikingly similar to each other for both enriched (Fig.S4B) and decreased genes (Fig.S4C) as shown by heat map representations of their gene expression levels and by scatter plot comparison of complete datasets indicating our purification and library production approach are highly reproducible (Fig.S3D). We found that 209 genes out of the 236, reported in Zfin as expressed in neural crest cells by 24hpf (Bradford et al., 2011), were expressed in the nuclear samples at 2 FPKMs or higher (Fig.S4E,F). Gene set enrichment (GSE) analysis revealed the presence of neural crest-relevant pathways implicated in the formation of neural crest derivatives, such as Wnt, PDGF, TGFβ and Notch (Fig.S4D, red nodes). In particular, the largest node from the GSE analysis consisted of 294 Wnt pathway genes, in line with previous evidence for its major involvement in migratory crest (Dorsky et al., 1998) and its primary role in differentiation of pigment and sensory neuron lineages (Pavan and Raible, 2012). The reduced, but not absent, representation of general metabolic pathways such as the TCA cycle, de novo purine biosynthesis and glycolysis (Fig.S4D) confirmed the value of profiling of small, specifically defined cell populations. Notably, we found a significantly decreased representation of genes involved in neuronal differentiation (axon guidance, opioid prodynorphin and GABA-B receptor II signalling) (Fig.S4D, blue nodes). As neural cell-types are intimately mingled with migratory neural crest cells, this reduction indicates that we can cleanly dissociate targeted and non-targeted cell-types by our purification protocol.

**Intron quantifications:** Starting from Ensembl gene models, we incorporated intron positions in a custom GTF file and quantified total read count for all introns and exons of a gene model, respectively. We reduced intron positions by 10% of total intron length from 5’ and 3’ ends to account for mis-annotated splice sites. Only introns located within genes whose exonic sequence was expressed at > 1 FPKM in nuclear samples were selected for analysis. Moreover, genes whose introns contain another gene or transcript were excluded. After differential expression, analysis comparing nuclear and polysomal beta-actin samples at 16ss, genome wide additive expression profile of all differentially enriched introns larger than 30 kb were obtained using ngsplot (Shen et al., 2014).

**Global transcriptional patterns in multiple datasets:** Global distribution profiles and genebody plots were obtained using deepTools package profiler and heatmapper tools (Ramírez et al., 2014).

**Clustering of transcriptional patterns using k-means algorithm:** To characterise bidirectional transcription at active gene promoters in nuclear sox10 datasets, we employed k-means clustering using seqMINER (Ye et al.,...
The clustering was performed using linear normalization, k=10 clusters and a window of ±1.5 kb from the active TSS set (ATAC_TSS). This procedure was applied to read associated with each strand of the sox10 nuclear, bactin nuclear and sox10 polysomal RNA-seq datasets. Corresponding heat maps were generated using seqMINER. K-means clustering was applied in a similar strand-specific fashion, using linear enrichment parameter, to identify and classify enhancers active specifically in neural crest cell nuclei, but using the active enhancer set (ATAC_enhancer). For cis-regulatory signature analysis, clusters 1 and 2 enriched in bidirectionally transcribed elements were selected and used to quantify specific bidirectional cis-regulatory modules transcription in neural crest (sox10 samples).

**Scatterplot quantification of transcribed elements:** Count tables for each k-cluster identified from ATAC_TSS and ATAC_enhancer datasets were generated using subread featureCount (Liao et al. 2014) for every two samples being compared (e.g. sample A and B). To generate dot plots in R, counts were normalized to reflect difference in sequencing depth between the two samples by dividing each count value of the sample with a greater number of reads (e.g. sample A) to a “downsampling factor” (no. of sample A reads/no. of sample B reads). To plot and obtain Pearson correlation coefficient values of ATAC_enhancer k-clusters whose transcription is more enriched in sox10 nuclear datasets, non-strand specific count tables for all clusters were generated and normalized for sox10 nuclear and bactin nuclear samples. Cluster 4 containing non-transcribed elements was excluded. For ATAC_TSS k-clusters, strand-specific count tables were generated and normalized to highlight k-clusters whose transcription is more enriched in sox10 nuclear datasets, according to negative or positive strand. To quantify the proportion of TSSs that are bidirectionally transcribed in sox10 nuclear compared to bactin nuclear, the number of elements that have a normalized count value of >0 on both strands were determined. R package dplyr “anti_join” function was used to identify bidirectional elements within a given cluster that are found only in sox10 nuclear or bactin nuclear.

**Ranking neural crest (NC)-specific cis-regulatory modules (CRMs):** To rank specific NC cis-regulatory modules from cluster 1 and 2, sox10 nuclear replicates and bactin replicate BAM files were merged and read counts for 11,655 elements ATAC_enhancer features were obtained using subread featureCount (Liao et al. 2014) and features expressed at FPKM>1 in bactin and sox10 samples were considered. The ratio of FPKM values between sox10 nuclear and bactin nuclear is termed fold-change (FC) for 11,655 non-null elements from Clusters 1 and 2. These CRMs were ranked according to FC value. CRMs were assigned to the proximal genes targets based on distance in whose regulation they are putatively involved. Using bedtools, a total of 11,655 CRMs were assigned to 4,767 genes. To quantify the additive effect of multiple enhancers on a single locus, we computed the Additive Fold Change (AFC), as a sum of FCs of all active CRMs assigned to a given locus. We ranked all loci according to their AFC values to identify a critical set of highly regulated loci, defined as genes whose relative specific enhancer transcription (measured by AFC) falls beyond this inflexion point. We assumed that these genes constitute neural crest transcriptional signature at 16-18ss.

**GREAT Analysis:** To annotate and assign biological significance to the identified CRMs with intermediate to high FC value (1<FC<5), bidirectionally transcribed specifically in the neural crest nuclei, we applied the GREAT tool (Genomic Regions Enrichment of Annotations Tool) (Hiller et al. 2013; McLean et al. 2010), which allows prediction of functional cis-regulatory regions by analysing the annotations of the genes lying proximal to them. The analysis was performed using GREAT default parameters: basal regulatory region extending 5 kb upstream and 1 kb downstream from TSS, with maximal 1 Mb extension and using the whole genome as a reference. We retained results obtained with a with a p<0.001 according to both binomial and hypergeometric tests.

**Repeated elements quantification:** Repeat analysis was carried out as previously described (Goke et al. 2015). Briefly, coordinates for repeats were downloaded from UCSC Genome Browser (v. Jan 19 2011) for danRer7 (Zv9). Read count for repeat positions were obtained using featureCount and FPKM values calculated accordingly (Liao et al. 2014). Differential expression was carried out using the rank product non-parametric method (Breitling et al. 2004) using the R package RankProd, with pfp<0.05 (Hong et al. 2006).
References

G. Abe, M. L. Suster, and K. Kawakami. Tol2-mediated transgenesis, gene trapping, enhancer trapping, and the Gal4-UAS system. *Methods in Cell Biology*, 104:23–49, 2011.

S. Anders, A. Reyes, and W. Huber. Detecting differential usage of exons from RNA-seq data. *Genome Research*, 22:2008–2017, 2012.

A. Apolloni, I. A. Prior, M. Lindsay, R. G. Parton, and J. F. Hancock. H-ras but not K-ras traffics to the plasma membrane through the exocytic pathway. *Molecular and Cellular Biology*, 20:2475–2487, 2000.

O. Armant, M. Marz, R. Schmidt, M. Ferg, N. Diotel, R. Ertzer, J. C. Bryne, L. Yang, I. Baader, M. Reischl, and et al. Genome-wide, whole mount *in situ* analysis of transcriptional regulators in zebrafish embryos. *Developmental Biology*, 380:351–362, 2013.

Y. Bradford, T. Conlin, N. Dunn, D. Fashena, K. Frazer, D. G. Howe, J. Knight, P. Mani, R. Martin, S. A. Moxon, and et al. ZFIN: enhancements and updates to the zebrafish model organism database. *Nucleic Acids Research*, 39:D822–829, 2011.

R. Breitling, P. Armengaud, A. Amtmann, and P. Herzyk. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS letters*, 573:83–92, 2004.

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

J. D. Buenrostro, B. Wu, H. Y. Chang, and W. J. Greenleaf. ATAC-seq: A method for assaying chromatin accessibility genome-wide. In F. M. A. et al., editor, *Current Protocols in Molecular Biology 109*, pages 21–29, 2015.

J. Bussmann and S. Schulte-Merker. Rapid BAC selection for tol2-mediated transgenesis in zebrafish. *Development*, 138:4327–4332, 2011.

T. J. Carney, K. A. Dutton, E. Greenhill, M. Delfino-Machin, P. Dufourcq, P. Blader, and R. N. Kelsh. A direct role for Sox10 in specification of neural crest-derived sensory neurons. *Development*, 133:4619–4630, 2006.

R. B. Deal and S. Henikoff. The INTACT method for cell type-specific gene expression and chromatin profiling in *Arabidopsis thaliana*. *Nature Protocols*, 6:56–68, 2011.

A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, 29:15–21, 2013.

R. I. Dorsky, R. T. Moon, and D. W. Raible. Control of neural crest cell fate by the Wnt signalling pathway. *Nature*, 396:370–373, 1998.

S. Driegen, R. Ferreira, A. van Zon, J. Strouboulis, M. Jaegle, F. Grosveld, S. Philipsen, and D. Meijer. A generic tool for biotinylation of tagged proteins in transgenic mice. *Transgenic Research*, 14:477–482, 2005.

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

M. Heiman, R. Kulicke, R. J. Fenster, P. Greengard, and N. Heintz. Cell type-specific mRNA purification by translating ribosome affinity purification (TRAP). *Nature Protocols*, 9:1282–1291, 2014.

S. Higashijima, H. Okamoto, N. Ueno, Y. Hotta, and G. Eguchi. High-frequency generation of transgenic zebrafish which reliably express GFP in whole muscles or the whole body by using promoters of zebrafish origin. *Developmental Biology*, 192:289–299, 1997.
R. C. Tryon, N. Pisat, S. L. Johnson, and J. D. Dougherty. Development of Translating Ribosome Affinity Purification for zebrafish. *Genesis*, 51:187–192, 2013.

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

M. Westerfield. *The zebrafish book. A guide for the laboratory use of zebrafish (Danio rerio)*. University of Oregon Press, 4th edition, 2000.

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

D. Yu, H. M. Ellis, E. C. Lee, N. A. Jenkins, N. G. Copeland, and D. L. Court. An efficient recombination system for chromosome engineering in *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America*, 97:5978–5983, 2000.

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