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 sox10 biotagging transgenic and BAC drivers. Wide-field image of TgBAC(Sox10:BirA-mCherry)^ox104a (ncBirA(BAC)) (A). Tg(Sox10:BirA-membCherry)^ct706a (ncBirA) and Tg(Sox10:BirA-membCherry)^ct706b (ncBirA(b)) transgenes at 24hpf (B, C). The two alleles (B, C) of sox10 Biotagging transgenic exhibit expression in different neural crest derivatives that are included in the overall sox10 expression pattern, by 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 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 (βactin) 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 (βactin) 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 RanGap-Cerulean-Avi proteins localize similarly to the nucleus of all cells in the embryo. Schematic of Avi-RanGap effector construct with ubiquitin promoter (ubiq) upstream of N-terminal Avi tagged RanGap elements (J). Confocal image of hindbrain (K) and eye (L) of Tg(ubiq:Avi-RanGap) embryo (nucAvi(ubiq)). Scale bars: 20µm, except 50µm 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-tag and BirA (lanes 2 and 3) embryos in which BirA expression varies while “Avi-Rpl10” expression is high. Anti-GFP Western blot detects both cleaved and uncleaved Avi-Rpl10 that is 30 kDa and 55 kDa, respectively.
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)^{700a}, Tg(bactin:Avi-Cerulean-RanGap)^{700b} (myoBitA:nucAvi(bactin)) 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;nucAvi samples compared to whole embryo transcriptome, demonstrating the relationship between the p-value and log-fold change (red=enriched, blue=decreased in ncBirA;nucAvi 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 heatmap 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>
<thead>
<tr>
<th>Biotagging BirA driver</th>
<th>Allele</th>
<th>Abbreviation</th>
<th>Expression</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tg(bactin:BirA-membCherry)</td>
<td>ct709a</td>
<td>ubBirA</td>
<td>ubiquitous</td>
<td></td>
</tr>
<tr>
<td>Tg(bactin:NLs-BirA-membCherry)</td>
<td>ct710a</td>
<td>ubBirA(NLS)</td>
<td>ubiquitous</td>
<td></td>
</tr>
<tr>
<td>Tg(ubiq:NLs-BirA-Cherry)</td>
<td>ox114</td>
<td>ubBirA(ubiq-NLS)</td>
<td>ubiquitous</td>
<td></td>
</tr>
<tr>
<td>Tg(ubiq:BirA-Cherry)</td>
<td>ox115</td>
<td>ubBirA(ubiq)</td>
<td>ubiquitous</td>
<td></td>
</tr>
<tr>
<td>Tg(sox10:BirA-membCherry)</td>
<td>ct706a</td>
<td>ncBirA</td>
<td>cranial neural crest derivatives</td>
<td>female-exhibit ubiquitous maternal expression</td>
</tr>
<tr>
<td>Tg(sox10:BirA-membCherry)</td>
<td>ct706b</td>
<td>ncBirA(b)</td>
<td>cranial and trunk neural crest derivatives</td>
<td></td>
</tr>
<tr>
<td>Tg(zic2a:NLs-BirA-membCherry)</td>
<td>ct708a</td>
<td>hbBirA</td>
<td>presumptive telencephalon and diencephalon, hindbrain primary neurons</td>
<td>expression gone by 2dpf</td>
</tr>
<tr>
<td>Tg(myl7:BirA-membCherry)</td>
<td>ct704a</td>
<td>myoBirA</td>
<td>myocardium</td>
<td></td>
</tr>
<tr>
<td>Tg(myl7:NLs-BirA-membCherry)</td>
<td>ct705a</td>
<td>myoBirA(NLS)</td>
<td>myocardium</td>
<td></td>
</tr>
<tr>
<td>Tg(kdrl:BirA-membCherry)</td>
<td>ct703a</td>
<td>endoBirA</td>
<td>vascular endothelium</td>
<td></td>
</tr>
<tr>
<td>TgBAC(sox10:BirA-Cherry)</td>
<td>ox104a</td>
<td>ncBirA(BAC)</td>
<td>neural crest and derivatives</td>
<td>female may exhibit ubiquitous maternal expression</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Biotagging Avi effector</th>
<th>Allele</th>
<th>Expression</th>
<th>Note</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tg(bactin:Avi-Cerulean-RanGap)</td>
<td>ct700a</td>
<td>nucAvi(bact)</td>
<td>ubiquitous</td>
</tr>
<tr>
<td>Tg(bactin:RanGap-Cerulean-Avi)</td>
<td>ct701a</td>
<td>nucAvi(bact-Cterm)</td>
<td>ubiquitous</td>
</tr>
<tr>
<td>Tg(ubiq:Avi-Cerulean-RanGap)</td>
<td>ox113a/b</td>
<td>nucAvi(ubiq)</td>
<td>ubiquitous</td>
</tr>
<tr>
<td>Tg(bactin:Avi-Cerulean-Rpl10)</td>
<td>ox111</td>
<td>riboAvi(bact)</td>
<td>ubiquitous</td>
</tr>
<tr>
<td>Tg(ubiq:Avi-Cerulean-Rpl10)</td>
<td>ox112</td>
<td>riboAvi(ubiq)</td>
<td>ubiquitous</td>
</tr>
</tbody>
</table>
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/TMIX 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 (memCherry) by a short sequence encoding the ribosome-skipping peptide of *Thosea asigna* virus (2A, Fig.1B-D). The membrane localization signal for memCherry 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 memCherry 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 (memCherry) 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-mCherry)*ex1115* (ubBirA(ubiq)) (Higashijima et al., 1997; Mosimann et al., 2011) (Fig.1E). Test experiments, using homozygote *Tg*(bactin:BirA-2A-membCherry)*ct709a* (ubBirA) and *Tg*(ubiq:BirA-2A-mCherry)*ex1115* (ubBirA(ubiq)) 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 recombinering 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:

plIndigobac_itol2_fw: TTCTCTGTGTGGTCTGGAGAATGAAACATGGAAGTCCGAGCTCATCGCTCCCTGCTCGAGCCGGCCCAAGTG
plIndigobac_itol2_rev: CCGGCCAACACCCGCTGACGCGAACCCCTTGGCCGCGCATATTATGATCCTCTAGATCATCTAGATCT

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)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)lox111 (riboAvi(bact)) and Tg(ubiq:Avi-Cerulean-Rpl10)lox112 (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 (DynMag™-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 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.
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 resuspended 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 loosefitting 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 (NPB: 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 (1.5 x 10^7 beads) of M-280 Streptavidin-coated Dynabeads prepared for RNA procedures (see preparation of beads) with inhibitor (Roche, cat. no.05892791001)). To purify nuclei, the suspension was incubated with 250 µg (1.5 x 10^7 beads) of M-280 Streptavidin-coated Dynabeads prepared for RNA procedures (see preparation of beads) with inhibitor (Roche, cat. no.05892791001)). To purify nuclei, the suspension was incubated with 250 µg (1.5 x 10^7 beads) of M-280 Streptavidin-coated Dynabeads prepared for RNA procedures (see preparation of beads) with inhibitor (Roche, cat. no.05892791001)). To purify nuclei, the suspension was incubated with 250 µg (1.5 x 10^7 beads) of M-280 Streptavidin-coated Dynabeads prepared for RNA procedures (see preparation of beads) with inhibitor (Roche, cat. no.05892791001)). To purify nuclei, the suspension was incubated with 250 µg (1.5 x 10^7 beads) of M-280 Streptavidin-coated Dynabeads prepared for RNA procedures (see preparation of beads) with inhibitor (Roche, cat. no.05892791001)). To purify nuclei, the suspension was incubated with 250 µg (1.5 x 10^7 beads) of M-280 Streptavidin-coated Dynabeads prepared for RNA procedures (see preparation of beads) with inhibitor (Roche, cat. no.05892791001)). To purify nuclei, the suspension was incubated with 250 µg (1.5 x 10^7 beads) of M-280 Streptavidin-coated Dynabeads prepared for RNA procedures (see preparation of beads) with inhibitor (Roche, cat. no.05892791001)). To purify nuclei, the suspension was incubated with 250 µg (1.5 x 10^7 beads) of M-280 Streptavidin-coated Dynabeads prepared for RNA procedures (see preparation of beads) with inhibitor (Roche, cat. no.05892791001)). To purify nuclei, the suspension was incubated with 250 µg (1.5 x 10^7 beads) of M-280 Streptavidin-coated Dynabeads prepared for RNA procedures (see preparation of beads) with inhibitor (Roche, cat. no.05892791001)). To purify nuclei, the suspension was incubated with 250 µg (1.5 x 10^7 beads) of M-280 Streptavidin-coated Dynabeads prepared for RNA procedures (see preparation of beads) with inhibitor (Roche, cat. no.05892791001)). To purify nuclei, the suspension was incubated with 250 µg (1.5 x 10^7 beads) of M-280 Streptavidin-coated Dynabeads prepared for RNA procedures (see preparation of beads) with inhibitor (Roche, cat. no.05892791001)). To purify nuclei, the suspension was incubated with 250 µg (1.5 x 10^7 beads) of M-280 Streptavidin-coated Dynabeads prepared for RNA procedures (see preparation of beads) with inhibitor (Roche, cat. no.05892791001)). To purify nuclei, the suspension was incubated with 250 µg (1.5 x 10^7 beads) of M-280 Streptavidin-coated Dynabeads prepared for RNA procedures (see preparation of beads) with inhibitor (Roche, cat. no.05892791001)). To purify nuclei, the suspension was incubated with 250 µg (1.5 x 10^7 beads) of M-280 Streptavidin-coated Dynabeads prepared for RNA procedures (see preparation of beads) with inhibitor (Roche, cat. no.05892791001)). To purify nuclei, the suspension was incubated with 250 µg (1.5 x 10^7 beads) of M-280 Streptavidin-coated Dynabeads prepared for RNA procedures (see preparation of beads) with inhibitor (Roche, cat. no.05892791001)). To purify nuclei, the suspension was incubated with 250 µg (1.5 x 10^7 beads) of M-280 Streptavidin-coated Dynabeads prepared for RNA procedures (see preparation of beads) with inhibitor (Roche, cat. no.05892791001)). To purify nuclei, the suspension was incubated with 250 µg (1.5 x 10^7 beads) of M-280 Streptavidin-coated Dynabeads prepared for RNA procedures (see preparation of beads) with inhibitor (Roche, cat. no.05892791001)). To purify nuclei, the suspension was incubated with 250 µg (1.5 x 10^7 beads) of M-280 Streptavidin-coated Dynabeads prepared for RNA procedures (see preparation of beads) with inhibitor (Roche, cat. no.05892791001)). To purify nuclei, the suspension was incubated with 250 µg (1.5 x 10^7 beads) of M-280 Streptavidin-coated Dynabeads prepared for RNA procedures (see preparation of beads) with inhibitor (Roche, cat. no.05892791001)). To purify nuclei, the suspension was incubated with 250 µg (1.5 x 10^7 beads) of M-280 Streptavidin-coated Dynabeads prepared for RNA procedures (see preparation of beads) with inhibitor (Roche, cat. no.05892791001)).
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, RNAsin, RNaseOUT, SUPERaseIN, 100 µg mL⁻¹ cycloheximide and 1% IGEPLA 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-cherry) 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(myoBirA) and the Avi effector allele, Tg(bactin:Avi-Cerulean-RanGap) 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 (∆∆Ct) method [Livak and Schmittgen [2001]. Primers used for qRT-PCR are as follows: myl7 (myl7_fw: AGGGGGAAAACTGCTCAAAG and myl7_rev: TGATACTCCATCCCGGTTC), vmhc (vmhc_fw: TCGTCAGTCGTGAAAGGAGGTGC and vmhc_rev: GGCTCATGAGGAGGTGAA), 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-seq 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 (Epitest). 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) 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 --slocal 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...
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 v 1.4.5 (Liao et al., 2014) or htseq-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 (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 wasexpressed 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 heatmap tools (Ramirez 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., 2014).
ATAC Additive Fold Change (AFC), as a sum of FCs of all active CRMs assigned to a given locus. We ranked all loci using the R package RankProd, with Differential expression was carried out using the rank product non-parametric method (Zv9). Read count for repeat positions were obtained using featureCount and FPKM values calculated accordingly. 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.

**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


