A genome-wide assessment of the ancestral neural crest gene regulatory network

Dorit Hockman1,2, Vanessa Chong-Morrison1,6, Stephen A. Green3, Daria Gavriouchkina1,7, Ivan Candido-Ferreira1, Irving T.C. Ling1, Ruth M. Williams1, Chris T. Amemiya4, Jeramiah J. Smith5, Marianne E. Bronner3 & Tatjana Sauka-Spengler1*

The neural crest (NC) is an embryonic cell population that contributes to key vertebrate-specific features including the craniofacial skeleton and peripheral nervous system. Here we examine the transcriptional and epigenomic profiles of NC cells in the sea lamprey, in order to gain insight into the ancestral state of the NC gene regulatory network (GRN). Transcriptome analyses identify clusters of co-regulated genes during NC specification and migration that show high conservation across vertebrates but also identify transcription factors (TFs) and cell-adhesion molecules not previously implicated in NC migration. ATAC-seq analysis uncovers an ensemble of cis-regulatory elements, including enhancers of Tfap2b, SoxE1 and Hox-a2 validated in the embryo. Cross-species deployment of lamprey elements identifies the deep conservation of lamprey SoxE1 enhancer activity, mediating homologous expression in jawed vertebrates. Our data provide insight into the core GRN elements conserved to the base of the vertebrates and expose others that are unique to lampreys.
The neural crest (NC) is a migratory embryonic cell population that is unique to vertebrates. NC cells form in association with the developing central nervous system, which they delaminate from after undergoing an epithelial-to-mesenchymal transition (EMT). They subsequently migrate throughout the body to give rise to a plethora of derivative cell types. The advent of the NC with its contributions to numerous tissues and organs is thought to have played an essential role in the diversification of vertebrates. Elucidating how the genetic signals involved in NC specification were modified over the course of vertebrate evolution is key to understanding how this diverse assemblage evolved and expanded. This requires a detailed picture of how the NC GRN functioned in the vertebrate ancestor. To this end, the sea lamprey, a basal vertebrate, serves as a good model. Morphologically, these animals are considered living fossils with a body-plan that has remained consistent over at least the last 400 million years.

The current view of the NC GRN has been compiled from data generated in jawed vertebrates. By taking a candidate gene approach to compare lamprey and gnathostome TFs and signalling molecules, we previously showed that many key NC genes were conserved in expression and function between lamprey and jawed vertebrates. These results suggested that the basic NC GRN was already present at the base of vertebrates, although some key regulators were missing from the lamprey NC specifier module.

Recently, our understanding of the NC GRN in gnathostomes has been increased with the advent of next-generation sequencing techniques including RNA-seq, ChIP-seq and ATAC-seq. Progress in reconstructing the NC GRN of jawless vertebrates, however, has been limited owing to incomplete genomic information. Recently, a germline assembly for the sea lamprey, that unlike previous assemblies is not affected by DNA elimination and has an increased contiguity to near chromosome-scale resolution, has made it possible to interrogate the regulatory genome of this basal vertebrate with increased confidence.

Here, we explore the dynamics of gene expression and chromatin accessibility during cranial NC specification and migration in the sea lamprey. By comparing our genome-wide representation of the active lamprey NC transcriptome to that of jawed vertebrates, our analyses highlight the components of the NC GRN that are conserved and likely to be essential for NC specification. We analyse the chromatin accessibility in the NC cells of two lamprey species, and find that cross-species mapping highlights putative cis-regulatory elements. Importantly, we identify enhancer elements that drive expression in the lamprey NC, and provide evidence that regulation of a SoxE family gene is conserved between jawless and jawed vertebrates. By adapting high-throughput tools to the lamprey, our data provide insight into the ancestral state of the NC GRN.

Results

Dynamics of the developing NC transcriptome. We obtained cranial NC RNA-seq data by dissecting the dorsal neural tube (DNT) including premigratory, early-delaminating and/or late-delaminating NC cells at Tahara (T) stage T18, T20 and T21 (Fig. 1a), respectively. In sea lamprey embryos, NC cells reside within the neural folds, which converge at T18 to form a neural rod and fuse at T20, when the first signs of NC migration have been reported.

Reads were mapped to the sea lamprey germline genome assembly. A consensus transcriptome consisting of 120,207 transcripts at 72,171 genetic loci was assembled de novo from the mapped DNT data sets, combined with mapped RNA-seq data sets from whole heads and whole embryos at T20. 67,736 of the transcripts did not overlap with any annotated genes and thus represent candidate novel transcripts or transcribed transposable elements. The latter were not integrated in the current conservative gene model annotation that excluded repetitive elements. Principal component analysis (PCA) of DNT count data showed clear separation along principal component 1 (PC1), which accounted for 90% of the variance, reflecting the developmental stage of the tissue (Fig. 1b). PCA and regression analysis confirmed that the replicate data sets at each stage were highly correlated, demonstrating high reproducibility (Supplementary Fig. 1). Differential expression analysis between the T18 and T21 samples, which represent the neural tube tissue and associated premigratory and late-delaminating cranial NC, respectively, revealed 9106 differentially expressed genes (DESeq2, adjusted p value < 0.05). Of these, 5400 were enriched at T21, whereas 3706 were depleted (Fig. 1c). As expected, fewer genes were recovered as differentially expressed when T18 and T20 samples, or T20 and T21 samples were compared (Supplementary Fig. 2a).

We assessed the dynamics of signalling molecules and TFs expressed during NC specification making use of the germline genome annotation in which lamprey gene models were assigned to likely vertebrate homologues. As expected, several bona fide NC markers were enriched at T21 when compared with T18 (Fig. 1c, Supplementary Data 1), Wnt11, which has a role in establishing the neural plate border and is maintained in the DNT, was one of the most significantly enriched genes, as were Wnt3 and Wnt10 (Fig. 1c). In contrast, several Wnt homologues (Wnt5a/b, Wnt7a, Wnt8a) were depleted at T21, consistent with studies showing that Wnt expression is dynamically modulated during NC delamination and migration. Genes play a role in neural tube development, such as Pax6a and Bmp4, as well as NC specifier genes like SoxE genes (SoxE1 and SoxE2), FoxD3, Msx2, Tjap2A and Tjap2B were increased by at least twofold at T21 when compared with T18, whereas other genes including several Hox, Tbx and Gata TFs were depleted (Fig. 1c), analogous to previous observations in gnathostomes. Similarly, premigratory NC genes (e.g., SoxE1 and FoxD3) were upregulated at T20 when compared with T18 (Supplementary Data 2a; Supplementary Data 1). When T21 was compared with T20, late NC specification factors and factors expressed at the onset of migration (e.g., SoxE2, SoxE3, CadN) were enriched (Supplementary Fig. 2a; Supplementary Data 1). Both Ets1b and Twist1 were depleted at T21 when compared with T18 (Fig. 1c), confirming previous findings regarding their absence from lamprey migratory NC.

Weighted Gene Co-expression Network Analysis (WGCNA) revealed 12 gene clusters with significantly higher gene expression at T18, and 13 gene clusters with significantly higher gene expression at T21, mirroring the results from our differential expression analysis (Supplementary Data 2, Supplementary Fig. 3). This approach delineated patterned sets of all expressed genes in NC cells. Tbx6 and Wnt5a were placed in a cluster of 767 genes that showed a drop in expression from T18 to T20, and remained low at T21 (Fig. 1d). When T20 expressed in the results from our differential expression analysis (Supplementary Data 2, Supplementary Fig. 3). This approach delineated patterned sets of all expressed genes in NC cells. Tbx6 and Wnt5a were placed in a cluster of 767 genes that showed a drop in expression from T18 to T20, and remained low at T21 (Fig. 1d). When T20 expressed in the results from our differential expression analysis (Supplementary Data 2, Supplementary Fig. 3). This approach delineated patterned sets of all expressed genes in NC cells. Tbx6 and Wnt5a were placed in a cluster of 767 genes that showed a drop in expression from T18 to T20, and remained low at T21 (Fig. 1d). When T20 expressed in the results from our differential expression analysis (Supplementary Data 2, Supplementary Fig. 3). This approach delineated patterned sets of all expressed genes in NC cells. Tbx6 and Wnt5a were placed in a cluster of 767 genes that showed a drop in expression from T18 to T20, and remained low at T21 (Fig. 1d). When T20 expressed in the results from our differential expression analysis (Supplementary Data 2, Supplementary Fig. 3). This approach delineated patterned sets of all expressed genes in NC cells. Tbx6 and Wnt5a were placed in a cluster of 767 genes that showed a drop in expression from T18 to T20, and remained low at T21 (Fig. 1d). When T20 expressed in the results from our differential expression analysis (Supplementary Data 2, Supplementary Fig. 3). This approach delineated patterned sets of all expressed genes in NC cells. Tbx6 and Wnt5a were placed in a cluster of 767 genes that showed a drop in expression from T18 to T20, and remained low at T21 (Fig. 1d). When T20 expressed in the results from our differential expression analysis (Supplementary Data 2, Supplementary Fig. 3). This approach delineated patterned sets of all expressed genes in NC cells. Tbx6 and Wnt5a were placed in a cluster of 767 genes that showed a drop in expression from T18 to T20, and remained low at T21 (Fig. 1d). When T20 expressed in the results from our differential expression analysis (Supplementary Data 2, Supplementary Fig. 3).
Other putatively co-regulated TFs involved in NC migration were also placed in this cluster (Sox21 and Zeb2), as well as signalling receptors and ligands (ERBB4, Ednrb, Sema3D/4C/5B), secreted matrix remodelling enzymes (MMP13, ADAM10, ADAMTS1), collagens (Col3a1/4a1/4a6) and lamprey orthologues involved in organisation of the extracellular matrix (Prolyl 4-hydroxylase subunit alpha-1 [P4ha1], Fibulin [Fbln2] and Creb3l1) (Fig. 1g; Supplementary Data 2: cluster3). This cluster also featured
downstream effectors ensuring proper differentiation into NC derivatives, such as melanocytes (RAB32, Sox21), neurons (Nrcam, Atoh1, Neurofascin orthologues, LRRN2) and glia (Sox2, GFRA1, Cdksr1, APP orthologues) (Fig. 1g: Supplementary Data 2: cluster3).

Importantly, the two largest WGCNA clusters contained genes that have not previously been implicated in NC development, including genes coding for cell-adhesion molecules, such as desmocollins (Dsc1/2 and Sdk1), and extracellular matrix proteins, such as Vitrin (Fig. 1e, f; Supplementary Data 2). Many novel TFs, as well as those that play a role later in NC development, were also in these clusters (Fig. 1g: Supplementary Data 2). For example, Nmi (N-myc interactor), which interacts directly with Sox10 and inhibits canonical Wnt signalling in cancer, showed increased expression at T20, whereas EHF (Ets homologous factor, also known as Epithelial Specific Ets-3), proposed to have a role as a tumour-suppressor in prostate cancer, and oncogene in ovarian cancer, showed expressed elevated expression at T21. Fli1 and Sabt2, which are expressed in the developing branchial arch cartilage and mesenchyme, and Nfatc, which forms a complex with Sox10 during Schwann cell differentiation, were also elevated at T21 (Supplementary Data 2).

Gene Ontology (GO) analysis of genes enriched at T18, and thus depleted at T21, highlighted terms associated with the NC GRN signalling module, including Wnt and FGF signalling (Fig. 1i). Terms associated with early embryonic processes such as gastrulation, axial patterning and neural tube development were also enriched at this stage. Conversely, genes enriched at T20 or T21 when compared with T18 revealed an overrepresentation of terms associated with early NC specification, and the development of tissues that receive a NC contribution, such as cartilage, bone, neural tissue and the heart (Fig. 1i; Supplementary Fig. 2b). The biological process terms associated with NC cell migration were specifically enriched only at T20. Although the GO term ‘NC cell differentiation’ was enriched at all stages (Fig. 1i), the genes associated with this term at NC stages, T21 and T20 showed little overlap with those at T18 (Supplementary Fig. 2c; Supplementary Data 3), thus providing evidence for a dynamic progression in gene expression as NC development progresses. The presence of GO terms associated with kidney development at T20 and T21 is largely owing to enrichment in genes known to be highly expressed in the NC, but also have a role elsewhere in the embryo (e.g., Ctnnb1, Sox9 and Bmp4).

Finally, we characterised the expression of a number of genes that were enriched at T21 using whole-mount in situ hybridisation (Fig. 1f; Supplementary Fig. 2d). As expected, Bmp4 and Tja2b were clearly expressed in the DNT at T21, whereas Tja2b was maintained in the migratory NC at T23. Interestingly, Fli1 expression was also associated with the DNT at T21, indicating that this factor may have an early role in NC development, as well as being expressed in the branchial arches later on. Similarly, Zfhx4, which is enriched in the migratory NC in the chicken, was associated with the neural tube and branchial arches, suggesting a role in both early and later NC development. Col6a1 was most strongly expressed adjacent to the neural tube, surrounding the developing otic placode, where it likely supports NC cell migration. Sdk1 and Vitrin were expressed in the region coinciding with the neural tube and delaminating NC at T23. Sdk1 was also associated with the developing branchial arch mesenchyme at T23. Tead1, a TF that is involved in regulating cell proliferation during development and in cancer, was expressed in the neural tube as well as the migratory NC at T23. The zinc finger protein Bnc2, which is implicated in craniofacial development and adult pigmentation, was associated with the DNT and the developing branchial arch mesenchyme at T23. Sdk1, Vitrin and Bnc2 expression at T21 was possibly too low for detection by in situ hybridisation (normalised readcount of 50.8, 119.4 and 378.4, respectively), whereas Tead1 was expressed most strongly in the mid-brain region including the NC (Supplementary Fig. 2d).

Taken together, our RNA-seq data confirm, with a higher level of detail, previous findings that a large proportion of the NC GRN is conserved to the base of the vertebrates. Importantly, our analyses reveal factors, whose role in NC development and diversification warrants further investigation.

**Genome-wide assessment of chromatin accessibility.** We next sought to explore the regulatory connections between players in the NC GRN. To this end, it is essential to identify cis-regulatory elements that control gene expression in the NC. ATAC-seq reveals regions of accessible chromatin, and enables a genome-wide assessment of putative cis-regulatory elements. We analysed chromatin accessibility in lamprey cranial DNTs or whole heads at T20, T21 and T23 (Fig. 2a), which encompass the early-delaminating, late-delaminating and post-migratory NC. Mapped biological replicate ATAC-seq libraries were highly correlated (Supplementary Fig. 4) and insert size distribution showed a stereotypical ~150 bp periodicity (Supplementary Fig. 5a), consistent with the nucleosome occupancy of chromatin.

Peak detection and annotation, using our de novo assembled transcriptome, was consistent across stages, with the majority of peaks found in intergenic and intronic regions where cis-regulatory elements are expected to be located (Fig. 2b). To focus our analyses on peaks associated with NC GRN genes, consensus peaksets for each annotation category (promoter, intergenic and intronic) were filtered to contain peaks that were associated with genes enriched at T21. Promoter peak groups were further filtered to contain elements associated with the promoters of genes annotated in the germline genome assembly.
K-means clustering of the ATAC-seq signal over consensus peaksets (8998 intergenic; 17,908 intronic; 1860 promoter) revealed the dynamics of chromatin accessibility genome-wide over the course of development (Fig. 2c–f). The ATAC-seq signal associated with promoter peaks was highest at T21 for all clusters showing, as expected, that enriched gene expression correlated with increased promoter accessibility (Fig. 2c). In addition, when all promoter peaks were taken into consideration (i.e., 10,286 annotated and novel promoter peaks), gene expression associated with promoter peaks was higher and less variable than that.

---

**Notes:**
- Fig. 2c–f: Illustrations showing the distribution and coverage of ATAC-seq peaks across different developmental stages (T20, T21, T23).
- K-means clustering: A method for grouping data points into clusters based on similarity.
- Chromatin accessibility: The openness of chromatin for transcriptional activity.
- Gene expression: The level of gene transcription in a cell.

---

**Diagrams:****
- Diagram a: Illustration of developmental stages at T20, T21, and T23 with respective replicates.
- Diagram b: Bar graph showing the number of peaks (thousands) across developmental stages.
- Diagram c: Graph depicting the mean read coverage at promoter peaks.
- Diagram d: Heatmaps illustrating the expression patterns across developmental stages.
- Diagram e: Heatmaps showing the expression levels of genes across clusters.
- Diagram f: Heatmaps highlighting the expression patterns of specific clusters.
- Diagram g: Heatmaps depicting the expression levels across different developmental stages.
- Diagram h: Heatmaps illustrating the expression patterns of specific clusters.
- Diagram i: Heatmaps showing the expression patterns across different clusters.
- Diagram j: Heatmaps depicting the expression levels across developmental stages.
associated with the intergenic and intronic peak clusters (Fig. 2e, Supplementary Fig. 5b).

We were particularly interested in cis-regulatory elements that regulate gene expression during EMT. K-means clustering of intergenic and intronic peaks revealed two large clusters (intergenic cluster 5; intronic cluster 4) that displayed increased accessibility at T21 and T23 compared to T20 (Fig. 2d, f). Gene ontology terms associated with intergenic cluster 5 included ‘regulation of localisation’, ‘positive regulation of cell-substrate adhesion’ and ‘regulation of cell motility’, whereas terms associated with intronic cluster 4 included ‘cell–cell junction organisation’, ‘cytoskeleton reorganisation’ and ‘positive regulation of cell migration’ (Supplementary Fig. 5c). In addition, these clusters contained elements associated with known NC GRN TFs, SoxE1 (intergenic cluster 5) and Tfp2B (intronic cluster 4).

To quantify the significance of these EMT clusters, we plotted the ATAC-seq signal levels of our peaksets at T20 against those at T23 and calculated the Pearson correlation coefficient for all intergenic and intronic clusters (Supplementary Fig. 5d). Both EMT clusters were significantly offset from all other identified groups of accessible elements, suggesting that the dynamic opening of putative cis-regulatory elements may single them out as functional enhancers during.

We used TF-binding site motif analysis to further interrogate the ATAC-seq k-means clusters. Similar to our recent ATAC-seq analyses in zebrafish11 and chicken12 NC, several intergenic clusters, including cluster 5, were enriched for Sox, Tfp2 and Tead motifs (Fig. 2g, h). GRHL2, Pax, Ets1 and Fox motifs were also enriched. Intronic cluster 4 displayed similar enrichment for Sox and Pax motifs, as well as Fox sites, whereas GRHL2 and REST-NR5F motifs showed the highest enrichment in this cluster (Fig. 2g, h). The presence of motifs for key NC specification TFs (Sox, Fox, Tfp, Ets1, Pax families6), as well those that play a role in EMT (GRHL228), suggested that these clusters harbour cis-regulatory elements that provide connections between NC GRN players. Enrichment of CTCF-binding sites in all intergenic clusters further suggests these peaks may represent putative cis-regulatory elements.

Cluster 1 for intergenic and intronic peaksets had a distinct TF-binding site profile from the other clusters, which resembled the binding profile of our promoter peakset (annotated and novel promoter peaks), and was enriched for motifs found in the HOMER promoter motif library (Fig. 2). These clusters consisted of peaks with a broad, open profile at all stages (see Fig. 2d). Therefore, it is likely that cluster 1 for both intergenic and intronic peaksets represent peaks with promoter-like activity.

Combinatorial TF binding at enhancers has been suggested to have an important role in NC GRN function11. We used in silico two-way combinatorial analysis to test for evidence of multiple TF co-activity at putative cis-regulatory elements. Focussing on intergenic cluster 5, we selected 18 TFs that displayed enriched motifs and enriched gene expression at T21 (Supplementary Data 4). The motifs of six TFs (Smd2, Sox10, TFAP2a, Pit1, Bcl6 and Pbx3) were significantly enriched in pair-wise analyses (*p < 0.05; two-tailed Chi-squared test) (Fig. 2i). Combinations of Sox10 sites were enriched, whereas Smd2 sites were enriched in combination with all the other five motifs. Interestingly, we found similar co-binding of Smd proteins with canonical NC TFs, such as Sox and TFAP2 proteins in the early chick NC12. This suggests that the combinatorial activity of TGFβ-signalling with canonical NC specification factors at enhancers is a conserved property of the NC GRN.

Active NC-specific cis-regulatory elements. To identify active NC-specific cis-regulatory elements, we used peaks from our EMT clusters that were associated with loci of known NC GRN genes in enhancer-reporter assays. An element, located 16.6 kb downstream of the SoxE1 locus (Fig. 3a), drove specific reporter expression in the delaminating NC cells from T21 and labelled the cells as they migrated into the branchial arches and contributed to known NC-derived structures, such as the branchial arch cartilage (Fig. 3b). Analysis of reporter expression alongside endogenous SoxE1 expression revealed an overlap in the delaminating NC at T23 and in the branchial arch cartilage at T26 (Fig. 3c). Together, these results suggest that we have identified a lamprey enhancer for SoxE1, which is activated in the delaminating NC. An element located in the third intron of the Tfp2B gene drove reporter expression in the migrating NC from T23 and labelled NC derivatives at later stages (Fig. 3d, e).

These cis-regulatory elements overlapped with peaks in similar ATAC-seq data sets from the brook lamprey (Lampetra planeri) (Fig. 3a, d). This surprising finding suggests conservation of cis-regulatory elements across lamprey species, which were separated at least 40 MYA37. Our analysis thus suggests a high degree of sequence conservation at the level of the functional non-coding regions of the genome of these two species, thus facilitating the identification of cis-regulatory elements using cross-species whole-genome alignment of ATAC-seq data.

Putative lncRNAs associated with lamprey Hox-a2 enhancers. Our ATAC-seq data set can be used to refine known cis-regulatory regions into modules that drive expression in specific tissues, including the NC. Sea lamprey cis-regulatory elements for Hox-a2 that drive gene expression in the neural tube, somites and NC are found in a 9-kb region upstream of Hox-a2, whereas elements that drive expression in the NC and somites alone are located within 4-kb of the Hox-a2 locus39. Recently, a core 1.5-kb region within the 4-kb enhancer (elementA) was shown to recapitulate the activity of the full 4-kb element39. These findings are confirmed by our ATAC-seq data set. We found that a ~1.5 kb element encompassing an ATAC-seq positive region at ~ ~ 8.5 kb drive...
reporter expression that was restricted to the neural tube (Fig. 4a, b: element-8.5 kb). In addition, a clear ATAC-seq-positive region was present over the 1.5 kb core NC enhancer element (Fig. 4a: elementA).

Interestingly, our RNA-seq data revealed bidirectional transcription from these loci, a known occurrence at active enhancers\textsuperscript{40} (Fig. 4a, bottom panel). Two novel transcripts from our transcriptome overlapped these regions: a 12,770 bp sense transcript (Fig. 4a, maroon label) and a 4206 bp, spliced antisense transcript (Fig. 4a, blue label). The longer sense transcript resembles the multiexonic enhancer (me)RNA transcripts reported in association with the ethryroid-specific intergenic...
**Fig. 3** Tissue-specific enhancer activity in the lamprey NC. **a, d** The SoxE1 **a** and Tfap2B **d** loci of the sea lamprey germline genome, with merged replicate ATAC-seq coverage tracks from the sea lamprey (Pm) and brook lamprey (Lp) at each developmental stage. Bars below coverage plots indicate peak regions identified with Macs2. The black box indicates the region tested in enhancer-reporter assays. **b** GFP reporter expression in lamprey embryos injected with the SoxE1 enhancer-reporter construct at the one-cell stage and allowed to grow to indicated stages (observed in 195 out of 1337 injected embryos). In transverse section (row 2, panels 2–6) GFP+ cells are visible delaminating from the neural tube at T22 (n = 2), migrating between the neural tube and otic vesicle at T23 (n = 2) and contributing to the branchial arch cartilage at T26 (n = 3). Coloured stars indicate panels showing the same embryo at successive developmental stages. Dashed boxed regions indicate regions magnified in adjacent panels. **c** Overlap of GFP reporter expression with native SoxE1 expression (magenta) in the delaminating and migrating NC at T23L (arrowheads; n = 4) and in the branchial arch cartilage at T25 (n = 2). **e** GFP reporter expression in lamprey embryos injected with the Tfap2B enhancer-reporter construct at one-cell stage and allowed to grow to indicated stages (observed in 25 out of 340 injected embryos). In transverse section at T26 (panel 5), GFP+ cells are visible in the branchial arch cartilage (n = 3).

Bac, branchial arch cartilage; L, late; Nt, neural tube; Obc, orobranchial cavity; Ov, otic vesicle. Scale bars for wholemounts: 200 μm. Scale bars for sections: 25 μm.

---

**Fig. 4** Characterisation of a Hox-a2 enhancer and associated transcription. **a** The Hox-a3/Hox-a2 locus in the sea lamprey germline genome, with merged replicate ATAC-seq coverage tracks from the sea lamprey (Pm) and brook lamprey (Lp) at each developmental stage, as well as RNA-seq coverage tracks from a representative T21 sample indicating directional transcription. Bars below ATAC-seq coverage plots indicate peak regions identified with Macs2. The black boxes highlight two ATAC-seq positive regions within the 9-kb region upstream of the Hox-a2 locus and the dashed boxes highlight bidirectional transcription over these regions. De novo assembled transcripts for the Hox-a2 locus are shown maroon (sense) and dark blue (antisense). **b** GFP reporter expression in lamprey embryos injected with the element-8.5 kb enhancer-reporter construct at one-cell stage and allowed to grow to indicated stages (observed in 10 out of 149 injected embryos). GFP reporter expression is seen in the neural tube. Scale bars: 200 μm.
enhancer, R4 in mice\textsuperscript{41}, as well as with multiple late NC-specific enhancers in zebrafish\textsuperscript{10}. As is the case with the R4 mRNA, the Hox-a2\texttext{-}8.5 kb upstream enhancer appears to initiate the transcription of an alternative first exon, which is spliced onto an adjacent annotated exon and reads through the remaining exons of the Hox-a2 gene. This results in a spliced transcript reminiscent of the annotated version, albeit with an extended 5'-UTR or first exon. Therefore the enhancer, located \textapprox 8.5 kb upstream of the Hox-a2 locus, may be acting as an alternative promoter (indeed, the peak was annotated as a promoter in our upstream of the Hox-

brain, heart, kidney and gonad RNA-seq datasets\textsuperscript{15}, whereas 70% were associated with ATAC-seq positive regions. The gnathos-
tome HoxA locus is known to harbour IncRNAs, including HOTAIRM\textsuperscript{42} and HOTTIP\textsuperscript{43}, which have been shown to modulate gene expression in cis. The putative lncRNA, identified between Hox-a3 and Hox-a2, is significantly enriched in the DNT at T21 when compared to T18 (threefold change; \textit{p.adj.} = 2.9E-49), suggesting it may regulate Hox-a expression in the neural tube and/or NC at T21.

**Lamprey SoxE1 enhancer activity is conserved in gnathos-
tomes.** Our study seeks to define the core components of the NC GRN that are conserved across vertebrates. This includes assess-
ing whether the activity of NC enhancer elements present in a basal jawless vertebrate is conserved in jawed vertebrates, despite 500 million years of independent evolution\textsuperscript{37}. We generated transgenic zebrafish carrying the lamprey SoxE1 enhancer upstream of a minimal promoter and GFP using the Activator (Ac)/Disassociation (Ds) (Ac/Ds) transposition system\textsuperscript{14}, which facilitates highly efficient transgenesis in zebrafish\textsuperscript{45}, and resulted in seven independent integrations of the SoxE1 enhancer:GFP cassette. Although only weak reporter expression was visible in F\textsubscript{0} embryos, the F\textsubscript{1} generation displayed striking heterospecific reporter expression in the branchial arches by \textapprox 30 hpf, mirroring the enhancer activity in the lamprey at T23 (Fig. 5a, Supplementary Fig. 6a, b). Later (\textapprox 60 hpf), reporter expression was visible in head structures that receive NC contributions including the branchial arch mesenchyme, cranial ganglia (Fig. 5; Supplementary Fig. 6c\texttext{-}e), as well as in melanocytes, and putative Schwann cells in the trunk (Fig. 5 row 2).

To further investigate the extent to which the lamprey SoxE1 enhancer activity is conserved among the gnathostomes, we used bilateral neural tube electroporation to introduce a lamprey SoxE1 enhancer:citrine reporter construct into the developing chicken NC at Hamburger Hamilton stage(HH)\textsuperscript{8}. At HH18, citrine enhancer activity is conserved among the gnathostomes, we used

DNT at T21 when compared to T18 (threefold change; \textit{p.adj.} = 8.77; Hox: 9.08). To assess the conservation of these putative binding sites we compared the sea lamprey genome sequence in the region of the SoxE1 enhancer to the same region in the juvenile brook lamprey and our brook lamprey ATAC-seq data (Supplementary Fig. 7). The majority of the binding sites were conserved across species, with only point mutations in two putative SoxE sites and one putative Hox-binding site (Supplementary Fig. 7). The high level of sequence conservation in this region supports a functional role for this enhancer across lamprey species and the presence of conserved TF-binding motifs suggests these sites may have a role in mediating cross-species enhancer activity.

To test whether cross-species enhancer activity is mediated by combinatorial TF activity in vivo, we used CRISPR/Cas9 to knockout the endogenous expression of selected TFs (hoxa2b, hoxb2a, hoxb3a, tfap2a, sox10) in zebrafish embryos, resulting from crossing our lamprey SoxE1 enhancer transgenic line reporting EGFP to a transgenic line expressing DsRed in the branchial arches (Fig. 5d\texttext{-}e). When DsRed-positive cells were sorted from the surrounding tissue in experimental embryos (Cas9 injected with target guide RNAs) and control embryos (Cas9-only injected), we found a significant downregulation of EGFP expression in the branchial arch cells upon targeted TF knockout as compared with controls (Fig. 5d, e; Supplementary Fig. 8). These perturbations support the notion that one or a combination of these short TF-binding site motifs is able to convey enhancer function\textsuperscript{47} and mediate the regulatory activity across evolutionary time, despite the lack of extensive sequence conservation.

Overall, our analyses indicate that the regulation of SoxE gene expression in the migratory NC is conserved to the base of the vertebrates and that enhancers with such conserved activity reflect the central lynchpin mediating the conservation of the NC GRN. Importantly, this suggests that regulation of one of the central players in the NC GRN has remained constant as the existence of the last common ancestor between jawed and jawless vertebrae.

**Discussion**

We present the most complete assembly to date of the lamprey NC GRN that can be directly compared with that of gnathos-
tomes. Our data enable identification of tissue-specific enhancers whose activity is evolutionary conserved and reveals putative non-coding RNA species.

Analysis of the lamprey NC transcriptional network provides global insight into the evolution of NC transcriptional pro-
grames. At premigratory NC stages (T20) in the lamprey, we observed similar gene enrichments in categories equivalent to those observed in zebrafish\textsuperscript{13} and chicken\textsuperscript{12}. The lamprey pre-
migratory NC shows significant functional enrichment in cate-
gories associated with mesenchymal (smooth muscle, connective tissue, cartilage and bone development) and neuronal (axonogenesis, gliogenesis) NC derivative fates. As development progresses and bona fide NC cells begin to delaminate (T21), enrichment terms changed to those characterising NC and stem cell programmes as well as autonomic nervous system formation. Interestingly, the lamprey NC GRN lamprey differs from that of zebrafish in that it lacks the NC sub-programme involved in the specification of the enteric nervous system (ENS). This is con-
sistent with studies showing that the lamprey may lack vagal NC and that the lamprey ENS may have much later onset\textsuperscript{48}.

Analysis of co-expression clusters using WGCNA increases the resolution of the putative lamprey NC GRN\textsuperscript{21} and suggests links
Fig. 5 The activity of the lamprey SoxE1 enhancer is conserved in gnathostomes. a In a 36 hpf transgenic zebrafish GFP reporter expression is visible in the developing branchial arches and melanocytes (n = 3). At 60 hpf, GFP+ cells populate the branchial arch cartilage and cranial ganglia in the head, whereas GFP+ melanocytes, melanoblasts and a putative Schwann cell are visible in the trunk (row 2) (n = 6). b In a HH18 chicken embryo, GFP expression is present in the branchial arches and the dorsal root ganglia (Drg) alongside cherry expression driven by the chick Foxd3 NC2 enhancer (n = 15). Dashed boxed regions indicate regions magnified in adjacent panels. c Schematic of CRISPR/Cas9 TF knockout experiment in transgenic zebrafish. d qPCR results indicating the change in EGFP expression levels relative to bactin in the branchial arches after injection with either Cas9 mRNA alone or Cas9 mRNA together with sgRNAs against hoxb2a, hoxb3a, hoxa2b, sox10 and tfap2a. Error bars: SD. Cg, cranial ganglia; Drg, dorsal root ganglia; Mb, melanoblasts; Mc, melanocytes; Ov, otic vesicle; Sc, putative Schwann cell. Scale bars in (a) 50 μm. Scale bar in (b) 200 μm
between early specification factors and their downstream effectors. TFs and signalling molecules associated with the NC specification module (e.g., Pax3/7, FoxD3, Tja2A, SoxE1) are upregulated at T20 and co-regulated with genes associated with delamination, including cell adhesion and cytoskeletal factors. Later, at T21, TFs associated with NC migration (e.g., Sox21, Zeb2, Tja2B) are co-expressed with genes associated with active migration, including signalling receptors, their ligands and secreted matrix remodelling enzymes. Interestingly, these genes are co-expressed with genes involved in NC differentiation. Future analysis using this resource to investigate direct links between co-regulated genes will allow expansion of the global structure of the lamprey NC GRN.

Our analysis of chromatin dynamics in the lamprey NC enabled the identification of tissue-specific enhancers. Similar to studies from C. elegans, where ATAC-seq analysis of whole animals at early embryonic, larval and adult stages revealed dynamically regulated cis-regulatory regions with tissue-specific activity49, we have performed this analysis using dissected cell populations predominantly, but not exclusively, comprised of NC cells at three successive stages of lamprey development. Together, these studies show that the dynamic signature associated with changes in chromatin accessibility over time can be used to pinpoint putative tissue-specific regulatory regions. Interestingly, we show that heterospecific analysis of ATAC-seq data from different lamprey species can provide clues for the location of conserved cis-regulatory regions. Although the evolutionary distance between the sea lamprey and gnathostomes precludes identification of cis-regulatory elements based on sequence conservation, the ATAC-seq reads from brook lamprey, L. planeri, were successfully mapped cross-species to the genome of the sea lamprey, P. marinus. This suggests that the putative functional non-coding elements have been conserved between the two lamprey species over the last 40 million years37. Thus, mapping the brook lamprey ATAC-seq data to the sea lamprey genome has enabled identification of conserved genomic regions.

We show that a lamprey SoxE enhancer drives tissue-specific reporter expression in the zebrafish and chicken NC. Interestingly, experiments in the invertebrate chordate, amphioxus, where the entire amphioxus SoxE locus and flanking genes were integrated into the zebrafish genome, resulted in reporter expression in the neural tube and tail bud, but not in the NC50. This suggests that a NC enhancer for SoxE expression is not present in the vicinity of the SoxE locus in amphioxus. Together, these results support the hypothesis that the acquisition of novel enhancers in early vertebrates was critical for the evolution of the NC-specific gene expression. Gain-of-function cis-regulatory changes, such as the appearance of new TF binding sites, likely facilitated co-option of pre-existing gene batteries, including the pro-chondrocytic SoxE genes and other mesenchymal gene programs, into NC-like cells at the neural plate border2. Indeed, we show that the lamprey SoxE1 enhancer harbours putative binding site motifs for important NC TFs, including Tja2 and SoxE factors, that activate and maintain SoxE transcription in the chick51,52, zebrafish53 and lamprey51. A Fox site is also present and could possibly control the activity pattern of the enhancer confined to specific regions of cranial NC conserved across vertebrates38. By knocking out the TFs associated with these binding sites in our transgenic zebrafish carrying the lamprey enhancer, we provide evidence to support the hypothesis that either single or a combination of TF binding sites mediate the conserved enhancer activity in the NC.

Taken together, our results suggest that the evolution of key TF-binding motifs was central to NC GRN evolution and that conservation of TF binding at these sites is important for the conservation of the NC GRN. Previous analyses of cis-regulatory TF-binding patterns across gnathostomes showed that, whereas binding motifs are conserved across species, their alignment or positioning is not54. This suggests that the functionality of cis-regulatory regions is independent of sequence constraint. We provide functional evidence that this hypothesis holds true to the base of the vertebrates by revealing that a cis-regulatory sequence from an agnathan, that does not align to the gnathostome genome, is able to drive tissue-specific gene expression in two gnathostome species. This conserved regulatory grammar may have been maintained over evolutionary time through positive-selection promoted by a requirement for co-operative binding of diverse TFs to direct tissue-specific gene expression. Support for this hypothesis comes from a recent study that analysed the location of binding sites for four liver-essential TFs across five mammalian species and found that cis-regulatory modules that harbour sites for all four TFs were more likely to be conserved than those with fewer shared sites55. It is plausible that physical interactions between the TFs have a role in the evolution and selection of higher order combinatorial TF binding. Indeed, thermodynamic biophysical modelling shows that the evolution of a functional binding site can be accelerated by cooperativity between adjacent TFs56.

In summary, by taking advantage of our highly contiguous germline genome assembly14, we have presented a genome-wide representation of gene expression and chromatin dynamics during lamprey cranial NC development. A limitation of our approach, which relied on dissection, is that analysed samples included some dorsal neural and ectodermal tissues, and thus our data represent a mixed cell population rather than a pure NC one. However, by focussing our analyses on the genes and chromatin regions that are being dynamically modified during the analysed time-points we were able to extract the expected signature for the cranial NC cells, which unlike skin and neural cells within this mixed cell population, are highly metabolically active and transcriptionally dynamic, owing to a later onset of their specification in the embryo. Taken together, our analyses uncover critical components of the NC GRN that are shared across vertebrates, as well as expose new players whose further investigation will expand our current view of the genetics of NC development.

**Methods**

**Lamprey husbandry and embryo dissections.** Adult sea lamprey (Petromyzon marinus) were supplied by the US Fish and Wildlife Service and Department of the Interior. Embryos obtained by in vitro fertilisation, were grown to the desired stage in compliance with California Institute of Technology Institutional Animal Care and Use Committee protocol #1436. Brook lamprey (L. planeri) embryos and ammocoete larvae were collected from a shallow river in the New Forest National Park, United Kingdom, with permission from the Forestry Commission and maintained in filtered river water at 13–19 °C. Prior to dissection, embryos were dechorionated in 0.1 × Marc’s Modified Ringers buffer (MMR) in a dish lined with 1% agarose. For dissection, embryos were moved into a shallow well in the agarose. T18, T20 and T21 DNTs including premigratory, early-delaminating and/or late-delaminating NC cells were dissected from the head using an eye-lash knife. T20 and T23 heads were dissected using forceps.

**RNA extraction and library preparation.** RNA was extracted from groups of at least 30 dissected DNTs at each stage, as well as from whole heads (two groups of 20) and whole embryos (two groups of 10) at T20. Tissue was lysed in the Ambion RNAqueous Total RNA Isolation kit lysis buffer (AM1931), set on ice for 15 mins with occasional vortexing, flash frozen in liquid nitrogen and stored at −80 °C. RNA was extracted using the Ambion RNAqueous Micro Total RNA isolation kit and assessed using the Agilent Bioanalyzer. Sequencing libraries were prepared from 100 ng RNA per sample using the NEBNext Ultra Directional RNA Library Prep Kit for Illumina (E7420) in combination with the NEBNext Poly(A) mRNA Magnetic Isolation Module (E7490) and NEBNext High-Fidelity 2 × PCR Master Mix (M0451S). Libraries were indexed and enriched by 15 cycles of amplification. Library preparation was assessed using the Agilent TapeStation and libraries qualified by Qubit. The concentration of library pools was assessed with the KAPA Library Quantification Kit (KK4835). Multiplexed library pools were sequenced using paired-end 75–100 bp runs on the Illumina NextSeq500 platform.
For dnt libraries and on the illumina HiSeq2500 platform for T20 heads and embryos.

**ATAC and library preparation.** Groups of dissected tissue were collected into L-15 medium (LifeTech) with 10% fetal bovine serum at 19°C. Tissue was first dis- sociated by pipetting up and down in dispase (1.5 mg/ml in DMEM; 10 mM Hepes, pH7.5), followed by the addition of an equal volume of trypsin (0.05% Trypsin; 0.53 mM EDTA in HBSS) at room temperature for a total of up to 15 mins. Dis- sociated cells were passed over a 40-μm cell strainer into Hanks’s solution (1 × HBSS; 10 mM Hepes; 0.25% BSA) and centrifuged at 500 × g for 7 mins at room temperature. The supernatant was removed and the fresh Hanks’s solution applied. 50,000 cells were counted out and centrifuged for 5 minutes at 500 × g at 4°C and washed with cold 2/3 phosphate-buffered saline (PBS) by centrifugation for 5 minutes at 500 × g, 4°C. The cells were resuspended in 50 μl cold lysis buffer (10 mM Tris-HCl, pH7.4; 10 mM NaCl, 3 mM MgCl2, 0.1% Igepal) and centrifuged for 10 minutes at 4 °C. Cells were then resuspended in Igepal and centrifuged for 30 minutes at 37 °C. To stop tagmentation, EDTA was added to final concentration of 50 mM and the reaction was incubated at 50°C for 30 mins. Tagmented DNA was cleaned up using the Qiagen MinElute PCR purification kit (2804) and amplified using the NEB Q5 High-Fidelity 2 × Master Mix (M0492S) for 14 cycles. The amplified library was cleaned up using the Qiagen MinElute PCR purification kit (2804) and XP AMPure beads (Beckman Coulter A63880). Tagmentation efficiency was assessed using Agilent TapeStation and libraries quantified by Qubit. The concentration of ATAC library pools was assessed with the KAPA Library Quantification Kit (KK4835). Multiplexed library pools were sequenced using paired-end 40 bp runs on the illumina NextSeq500 platform. The high correlation of the mapped ATAC-seq signal between biological replicates at each stage (Pearson's R > 0.9) confirms the reproducibility of our experimental approach (Supplementary Fig. 5).

**Pre-processing of next-generation sequencing reads.** Read quality was eval- uated using FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were trimmed to remove low quality bases using Sickle (https://github.com/najoshi/sickle) using the parameters -I -q 20.

**RNA-seq analysis.** Reads were mapped to the sea lamprey germline genome assembly27 using STAR (v2.4.2)27 (STAR --genomeDir $GENOME --read-DnaN $R1.fastq $R2.fastq -runThreadN 4 –outDataNamePrefix $PREFIX –fixReadsCommand zcat –outSAMreadField intronMotif –alignEndsType EndToEnd –outUnmappedFastx –outSAMtype BAM SortedByCoordinate). Separate transcriptomes for DNT sample datasets or head and embryo sample data sets were assembled de novo with Cufflinks (v2.2.1)28 using Cuffmerge default parameters58 to make a consensus transcriptome from all the data sets. Read counts for DNT data sets were obtained with Subread featureCounts (v1.4.6-p4)59 using the Cuffmerge consensus transcriptome in SAF format as a reference (featu- reCounts -p -B -M -F SAF -s 2 -T4 -a $SAF -o $OUT $IN.bam). Differential expression analyses were run on the R platform (v3.2.1; http://www.R-project.org/). The differential expression analyses were performed using DESeq2 (v1.2.8)60, weighted correlation network analysis (WGCNA)22 was per- formed on the variance stabilised normalised gene count tables generated by DESeq2 (filtered to only contain genes with a normalised count > 8) according to the pipeline detailed in the online WGCNA tutorial (https://labs.genetics.ucsc.edu/horvath/CohesionNetwork/Rpackages/WGCNA/). A soft-thresholding power of 2 and minimum module size of 100 was used for WGCNA. Both of these analyses were run on the R platform (v3.2.1; http://www.R-project.org/).

The average normalised gene counts that were associated with ATAC-seq peaks were obtained for stage T21 samples (see ATAC-seq analysis) were plotted in R using ggplot219. Output transcript lists from the differential expression analysis and WGCNA were annotated using the gene models associated with the sea lamprey germline genome assembly. Hox and Sox genes shown in figures were manually annotated with lamprey-specific gene names. Heatmaps of the average variance stabilised normalised gene counts were generated in R using heatmap2. Gene Ontology (GO) analysis was performed on annotated differentially expressed gene sets using the genes with baseMean value > 20, log2 FoldChange > 1 for upregulated and < -1 for downregulated genes and p-adjusted value < 0.01.

**ATAC-seq analysis.** Reads were mapped to the sea lamprey germline genome assembly27 using Bowtie220 (bowtie2 -p phred33 -X -2000 --very-sensitive –X -1 -S $SAF) and filtered with BamTools26 to remove unpaired reads and reads mapping to the mitochondrial and rDNA genomes. Filtered bam files were sorted using Samtools24 and paired-end bed files were obtained using bedtools (v2.15.0)25 bamtoolify.bedpe. Reads were extended to a read length of 100 bp. Peak-calling was performed using MACS262 (macs2 callpeak -t $IN.bed -f BED -name $IN.macs2 --outdir $OUT --shiftsize=100 --nomodel --slocal 1000). Output peak files (xls) for each develop- mental stage were converted to bed format and merged with bedtools merge to create one consensus peakset. The consensus peakset was annotated with Homer (v4.7; http://homer.salk.edu/homer/nphs) annotatePeaks.pl using the Cuffmerge gene models (with genes less than 1500 bp in length removed) as a reference. Annotated peaks were separated into intergenic, intronic and promoter peaksets according to their Homer annotation. Promoter peak sets were filtered with bedtools flank to only include elements that overlapped a region of up to 2 kb upstream of the sea lamprey germline genome gene models (bedtools flank -i promoters.bed -g germline_genome $IN.macs2 chrom.sizes -b 20000 -a 0) – Intergenic peaks that overlapped with promoters annotated in the sea lamprey germline genome gene models (i.e. gene models that were not present in the de novo cuffmerge assembly) were identified with bedtools intersect and moved to the promoter peakset. The inter- genic and intronic peaksets were further filtered to only contain peaks that were < 10 kb away from genes that were enriched at stage T18 with T18 (see RNA-seq analysis). k-means clustering of ATAC-seq signal over the final peakset was carried out using SeqMinER63 (1,500 bp on each side; no auto-turning; wiggle step: 15; k-means enrichment linear). Read counts for the ATAC-seq signal were obtained with Subread featureCounts (v1.4.6-p4) using the peakset format as a reference. Raw SAF format was converted to bed format and reads matched to the sea lamprey germline genome gene models (bedtools setCluster.saf). Correlation analysis on ATAC-seq readcount data was performed in R using plot and cor (method='pearson'). GO analysis was performed on the differentially expressed genes associated with intergenic and intronic "EMT" clusters using the PANTHER Overrepresentation Test (v11)64 with complete GO term databases for Mus musculus. Output GO terms were filtered to only contain terms that were enriched by at least 1.8-fold. Remaining GO terms were sum- marised with REVIGO65 and subsequently filtered to only contain terms with -log10 p value > -1.5.

Motif analysis was performed on the intergenic and intronic peakset clusters with Homer (v4.7) foundMotif.pl with a random set of 2329 (average number of peaks across all intergenic and intronic clusters) non-coding genomic regions as control. Heatmaps were generated in R using ggPlot geom-tile. To assess predicted co-binding frequencies for 18 selected TF motifs (selected based on their expression enriched in intergenic cluster 5) the E1 enhancer cluster, Supplementary Data S2, all possible pairs of motif combinations were computed in R (v3.2.1), using intergenic cluster 6 as a control. HOMER (v.4.7) annotatePeaks.pl script was used to screen such motifs in windows of 2250 bp from the peak centre. Motif instances were then converted into a matrix of presence (1) or absence (0) of motif occurrences in individual genomic regions. This matrix was then used as input into Cufflinks co-occurrence script to calculate motif co-occurrence occurrences. Motif co-occurrence comparisons at a = 5% (two-tailed Chi-squared test) with Bonferroni correction for multiple hypothesis (m) testing were retained for P values < α/m. Co-binding predictions were plotted using the Circize package in R (v3.2.3). HOMER (v.4.7) annotatePeaks.pl with options -m and -mscore, was used to locate enriched motifs that were identified using foundMotif.pl in the full-length SoxE1 enhancer.

**L. planeri ATAC-seq data mapped to SoxE1 enhancer genomic region of the sea lamprey germline genome was extracted using samtools view. The consensus sequence for all stages was generated in IGV (v2.3.60).**

In **vivo enhancer-reporter assays in lamprey.** A region of 1.5 kb encompassing the ATAC-seq positive accessible chromatin region to be tested was amplified by PCR and subcloned into the sea lamprey germline genome using primers designed with SnapGene software (Clontech), cloned into the HLC GFP reporter vector66 by In-Fusion HD Cloning (Clontech) and sequenced. Isecl-1 meganuclease-mediated transgenesis38,69 was performed in sea lamprey embryos. At 2-6 hours post fertilisation, single-cell embryos were injected with the Isecl-1 vector digestion mix at 20 ng/μl and maintained at 18°C in 0.1 × MMR for the remainder of development. At 1 dpf embryos were transferred to 96-well plates until 6 dpf when they were returned to petri dishes, and screened daily for reporter expression. Live embryos were imaged on a depression slide using a Zeiss Scope.A1 microscope fitted with a Zeiss AxioCam MRm camera and Zeiss ZEN 2012 software (blue edition).
each in PBST. Embryos were treated with 20 μg/ml solution of proteasine K in PBST at room temperature, for precisely two minutes, then immediately rinsed twice in PBST. Samples were then washed five times in 0.1% Triethanolamine (TEA) a total of three times. Embryos were treated with 0.25% Acetic Anhydride in TEA for 12 mins, and then in 0.50% Acetic Anhydride for another 12 mins.

Embryos were washed 2 x 5 mins in PBST, then rinsed in 4% Formaldehyde (diluted from 37% Formaldehyde in PBST) and rinsed 3 x 5 mins or more in PBST. Embryos were then incubated 3 x 5–10 mins in Hybridisation Buffer (hyb), then incubated for 3 h at 68 °C in 500 μl of fresh hyb. This solution was replaced with fresh, hot hybridisation buffer preincubated with each appropriate antisense digoxigenin-labelled riboprobe diluted to ~ 1 ng/ml of hyb and vials were incubated at 68 °C overnight. After 12–16 h, hybridisation solution was removed and replaced with fresh hyb. Samples were washed 3 x 30 min in pre-heated hyb solution. Most hyb was removed, except for ~ 500 μl, and an equivalent volume of pre-heated 2 x SSC was added, lightly mixed, and vials were incubated at 68 °C for 15 mins. Samples were then washed 3 x 30 mins in 2 x SSC, followed by 3 x 30 mins in 0.2 x SSC. Glass vials were allowed to slowly equilibrate to room temperature, then embryos were rinsed several times in RT MABT, then rinsed into MABT block (1 x MAB plus 2% Roche blocking solution and 0.1% Tween). Embryos were incubated in 500 μl of fresh MABT block for 1 h at RT, and then blocking solution was replaced with antibody solution and left to incubate overnight at 4 °C with slow shaking. Vials were washed twice in MABT block, 30 min each, then washed 5 x 30 mins in MABT, and 2 x 60 mins or more in MABT. Embryos were removed from borosilicate vials and placed into 12-well culture plates, rinsed 3 x 10 min in alkaline phosphate buffer. Buffer was replaced with BM Purple and samples were kept dark. When colour reaction was complete, embryos were rinsed in MABT, then post-fixed in 4% paraformaldehyde overnight at 4 °C. Embryos were dehydrated in a graded series of 100% methanol 4–5 times, and cleared for 3 x 3–5 mins in Murray’s Clear (1:2 mix of benzyl alcohol and benzyl benzoate). Murray’s Clear was removed and replaced with Permount, and embryos were individually mounted on slides, with coverslips having clay feet to act as spacers. Embryos were visualised on a Zeiss AxioImager.

Lamprey cryosectioning and immunostaining. Embryos were fixed at 4 °C overnight in 4% paraformaldehyde in PBS. Fixed embryos were incubated in PBS with 5% sucrose for 4 h at room temperature, followed by incubation overnight at 4 °C in 15% sucrose in PBS. Embryos were transferred into pre-warmed 7.5% gelatine in 15% sucrose in PBS and incubated overnight at 37 °C, before being transferred to pre-warmed 20% gelatine in PBS. Embryos were embedded in rubber moulds and frozen by immersion in liquid nitrogen. Blocks were cryosectioned at 6–10 μm. Gelatine was removed from the slides by a 5-minute incubation in PBS pre-warmed to 37 °C. For Immunostaining, slides were incubated overnight at room temperature in Alexa 488 conjugated anti-GFP antibody (Rabbit, 1:250; Life Technologies; A-21211) in blocking solution (10% donkey serum or 3% BSA in PBT) for 2 h and washed in antibody solution (Rabbit anti-GFP, 1:200 in block, Torrey Pines, TP401; mouse (IgG2b) anti-Ell3b/Alu (HoCd), 1:250 in block), then incubated in PBST for 2 h at room temperature. After several PBST washes, embryos were imaged in whole-mount on a Zeiss LSM 780 upright multiphoton confocal microscope with Zeiss ZEN 2011 black edition). To obtain sections of whole-mount immunostained samples, embryos were cryoprotected in 30% sucrose in PBS and embedded in OCT by immersion in liquid nitrogen. Sections were imaged on a Zeiss LSM 880 inverted confocal microscope with Zeiss ZEN 2.3 (black edition).

Fluorescent in situ hybridisation chain reaction. For in situ hybridisation chain reaction (HCR)71, a kit containing a DNA probe set, a DNA HCR amplifier, and hybridisation, wash and amplification buffers were purchased from Molecular Instruments for each target mRNA. The GFP and Sox2E1 probes initiate B3 (Alexa-546) and B4 (Alexa-647) amplifiers, respectively. Embryos were fixed overnight at 4 °C in MEMFA. They were then rinsed in PBST, dehydrated in a graded series of ethanol washes and stored in 100% ethanol at –20 °C. Following dehydration, embryos were treated with 10 mg/ml Proteasine K for 2.5 min and post-fixed in 4% paraformaldehyde for 20 min. Embryos were incubated with the probes in hybridisation solution overnight at 37 °C and, following appropriate washes, incubated in the hairpin solution in amplification buffer overnight at room temperature protected from light. Embryos were then imaged in whole-mount on a Zeiss LSM 780 upright multiphoton confocal microscope with Zeiss ZEN 2011 (black edition).

Splinkerette PCR. For splinkerette analysis72 five positive F2 zebrafish embryos from a single F1 parent outcrossed to wild type were collected and genomic DNA extracted. Approximately 500 ng of genomic DNA was digested overnight with Alul in a 30 μl reaction. Digested genomic DNA was purified using phenol:chloroform followed by ethanol precipitation before ligation with annealed splinkerette adaptors (CGAATCGAAGGTCTCGAGAACATGTGACAGATTGTGCTCGTCTCCGAGACCTAGGTTAAT and ATAAATGGCCTGTTCCGCATACTTACATAG) using T4 DNA ligase at 16 °C overnight in a thermal cycler. The adaptor-ligated genomic DNA was purified using Zymo Research Clean & Concentr (Cat. # D4003) and 20 ng of purified product used in a primary PCR reaction. PCR was performed using the following primers: CGAATCGAAGGTCTCGAGAACATGTGACAGATTGTGCTCGTCTCCGAGACCTAGGTTAAT (binding to adaptor) and GTTCGCTCGGCAGAAATTTAG (binding to Ds-3 integration arm), with 63 °C annealing temperature and 3 min extension time. 1 μl of primary PCR reaction was then used in 50 μl nested PCR reaction using the following primers: TCAGTACGAGAATCTGCTGCTCCCTCG (binding to adaptor) and CGGGTAGAGGTATTATTTACCGAC (binding to Ds-3 integration arm), with 60 °C annealing temperature and 5 mins extension time. The nested PCR was run on agarose gel to visualise number of integrations.

Chicken embryo in ovo electroporation. The lamprey Sox2E1 enhancer was cloned into the pTκ vector73 with GFP replaced by citrine, using In-fusion cloning (Clontech). Fertilised wild-type chicken eggs were obtained from Henry Stewart & Co, disabled and electroporated in a -pTK-citrine and N2-pTK-mCherry46 constructs were injected into the neural tube at 3.9 μg/μl each and electroporated bilaterally with 3 + 3 50 μs pulses at 12.5 V with 100 ms rest between pulses. The embryos were incubated at 37 °C until HH18 and imaged in whole-mount using a Zeiss LSM 780 upright multiphoton confocal microscope with Zeiss ZEN 2011 (black edition).

CRISPR/Cas9 assessment of Sox2E1 upstream factors. Oligo templates for guide RNAs (sgRNAs) targeting splice acceptor/exon boundaries of functional domains for hoxb2a, hoxb3a, hexa2b, sox10 and gap2a were transcribed and RNA in vitro transcribed using HiScribe T7 Quick High Yield RNA Synthesis Kit (New England Biolabs, E2030), followed by purification using MEGAClear Transcription Clean-Up Kit (ThermoFisher Scientific). Three oligos per target gene (Supplementary Table 1) were microinjected with Cas9 mRNA and two sgRNAs into one-cell stage embryos obtained by crossing Tg(sox2E1:dDsRed2)pd37 with Tg(CAG-EGFP)ox164 and Tg(tcf21:DsRed2p67). For each condition (Cas9 mRNA only, Cas9 mRNA with sgRNAs), ~ 25–30 DsRed-positive cells that included branchial arches were obtained from 3 days post fertilisation (dpf) embryos by FACS on BD FACSAria Fusion. Total RNA was extracted and DNasel-treated according to manufacturer’s protocol using RNAsqueous-Micro Total RNA Isolation Kit (ThermoFisher, AM1319). Reverse transcription was performed on three biological replicates using GoScript Reverse Transscriptase and Oligo(dT) (Promega). cDNA was diluted in water and 1 μl of cDNA was used in qPCR reaction with two 3.0 μl of primary PCR reaction was then used in 50 μl nested PCR reaction using the following primers: TCAGTACGAGAATCTGCTGCTCCCTCG (binding to adaptor) and CGGGTAGAGGTATTATTTACCGAC (binding to Ds-3 integration arm), with 60 °C annealing temperature and 5 mins extension time. The nested PCR was run on agarose gel to visualise number of integrations.
References

1. Le Douarin, N. & Kalcheim, C. The neural crest. (Cambridge university press, 1999).
2. Green, S. A., Simes-Costa, M. & Bronner, M. E. Evolution of vertebrae as viewed from the crest. Nature 520, 474–482 (2015).
3. Gans, C. & Northcutt, R. G. Neural crest and the origin of vertebrates: a new view. Development 142, 242 (2015).
4. Sauka-Spengler, T. & Bronner-Fraser, M. Insights from a sea lamprey into the evolution of neural crest gene regulatory network. Dev. Biol. 214, 303–314 (2008).
5. Gess, R. W., Coates, M. I. & Rubidge, B. S. A lamprey from the Devonian.

ARTICLE NATURE COMMUNICATIONS | https://doi.org/10.1038/s41467-019-12687-4 | www.nature.com/naturecommunications

Received: 3 March 2018; Accepted: 23 September 2019; Published online: 16 October 2019

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

Data availability

The authors declare that all data supporting the findings of this study are available within the article and its supplementary information files or from the corresponding author upon reasonable request. The data sets generated during and/or analysed during the current study have been deposited in the NCBI GEOarchive database under accession code GSE112072 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE112072.

Code availability

The custom script used to calculate motif co-occurrences is available at https://github.com/usabichick/NC-GRN. Details of software versions and parameters used, when different from default, are indicated in the Methods section.
50. Jandzik, D. et al. Evolution of the new vertebrate head by co-option of an ancient chordate skeletal tissue. *Nature* 518, 534–537 (2015).
51. Murko, C. & Bronner, M. E. Tissue specific regulation of the chick Sox10E1 enhancer by different Sox family members. *Dev. Biol.* 422, 47–57 (2017).
52. Betancur, P., Sauka-Spengler, T. & Bronner, M. A. Sox10 enhancer common to the otic placode and neural crest is activated by tissue-specific paralogs. *Development* 138, 3689–3698 (2011).
53. Veenstra, A. et al. Tel2-mediated control of sox expression facilitated the evolutionary emergence of the neural crest. *Development* 139, 720–730 (2012).
54. Schmidt, D. et al. Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* 328, 1036–1040 (2010).
55. Ballester, B. et al. Multi-species, multi-transcription factor binding highlights context. Bio-control of tissue-specific biological pathways. *eLife* 3, e02626 (2014).
56. Tugrul, M., Paixao, T., Barton, N. H. & Tkacik, G. Dynamics of transcription factor binding site evolution. *PLoS. Genet.* 11, e1005639 (2015).
57. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29, 15–21 (2013).
58. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* 7, 562–578 (2012).
59. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30, 923–930 (2013).
60. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Ontology and Reactome pathways, and data analysis tool enhancements.*
61. Murko, C. & Bronner, M. E. Tissue specific regulation of the chick Sox10E1 enhancer by different Sox family members. *Dev. Biol.* 422, 47–57 (2017).
62. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842 (2010).
63. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Bioinformatics* 29, 902–903 (2013).
64. Ye, T. et al. seqMINER: an integrated ChIP-seq data interpretation platform. *Nucleic Acids Res.* 45, D183–D189 (2017).
65. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842 (2010).
66. Barnett, D. W., Garrison, E. K., Quinlan, A. R., Strömberg, M. P. & Marth, G. A. Novel VEP format for summarizing and visualizing long lists of gene ontology terms. *PLoS. ONE* 6, e12809 (2011).
67. Ye, T. et al. seqMINER: an integrated ChIP-seq data interpretation platform. *Nucleic Acids Res.* 39, e35–e35 (2010).
68. Supek, F., Bosl, M., Skunca, N. & Smuc, T. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS. ONE* 6, e21800 (2011).
69. Parker, H. J., Sauka-Spengler, T., Bronner, M. & Elgar, G. A reporter assay in lamprey embryos reveals both functional conservation and elaboration of vertebrate enhancers. *PLoS. ONE* 9, e85492 (2014).
70. Bell, T. & Kramvis, A. Fragment merger: an online tool to merge overlapping long sequence fragments. *Viruses* 5, 824 (2013).
71. Choi, H. M. T. et al. Third-generation in situ hybridization chain reaction: multiplexed, quantitative, sensitive, versatile, robust. *Development* 145, (2018).
72. Uren, A. G. et al. A high-throughput splinkerette-PCR method for the isolation and sequencing of retroviral insertion sites. *Nat. Protoc.* 4, 789–798 (2009).
73. Uchikawa, M., Ishida, Y., Takemoto, T., Kamachi, Y. & Kondoh, H. Functional analysis of chicken Sox2 enhancers highlights an array of diverse regulatory elements that are conserved in mammals. *Dev. Cell* 4, 509–519 (2003).
74. Jao, L.-E., Wente, S. R. & Chen, W. Efficient multiplex biallelic zebrafish genome editing using a CRISPR nuclease system. *Proc. Natl. Acad. Sci.* 110, 13904–13909 (2013).
75. Kikuchi, K. et al. epicardiocyte cell-cell contact regulates cardiac morphogenesis. *Development* 138, 2895–902 (2011).