Supporting Information

Royo et al. 10.1073/pnas.1109037108

SI Materials and Methods

***In Silico* Screening for Conserved Noncoding Sequences in Metazoans.** We obtained a list of 59 automatically predicted conserved noncoding regions (CNRs) conserved between amphioxus and vertebrates from ref. 1. We further curated this list manually and excluded the elements that did not fulfill strictly the following criteria: (*i*) There is a minimum sequence identity of 60% over at least 60 bp; (*ii*) the CNRs are linked to the same genes in the different groups; (*iii*) the CNRs are conserved in amphioxus and human and in at least one other vertebrate group [fishes (zebrafish, fugu), amphibians (frog), and/or birds (chicken)], (*iv*) the CNRs are not repetitive elements; and (*v*) the CNRs are not unannotated protein coding sequences. After applying these filters and combining adjoining CNRs, the list was reduced to 13 CNRs present at least in the last common ancestor of extant chordates (Table S1).

We then assessed whether these elements were conserved in other animal phyla. We surveyed the genomes of *Nematostella vectensis* Joint Genome Institute (JGI) v1.0, *Daphnia pulex* JGI v1.0, *Lottia gigantea* JGI v1.0, and *Capitella teleta* JGI v1.0 at the JGI webpage (http://genome.jgi-psf.org/euk_home.html); *Strongylocentrotus purpuratus* Build 2.1 at the National Center for Biotechnology Information (NCBI) webpage (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi); and *Saccoglossus kowalevskii* 2008-Dec-09 linear scaffolds at the Human Genome Sequencing Center, Baylor College of Medicine Blast web page (http://blast.hgsc.bcm.tmc.edu/blast.hgsc?organism=20). We used two complementary strategies: (*i*) We conducted BLASTN searches of the CNRs amphioxus and human sequences against whole-genome targets, and (*ii*) we identified the orthologous associated gene and obtained the adjoining 20 kb upstream or downstream (depending of the position of the element in chordates) for the different species and performed ClustalW alignments with available CNRs. All conserved sequences had >60% identity in ungapped alignments in >60 bp.

Orthology Relationships of *Nematostella* SoxB Genes. Previous studies (2, 3) could not identify clear orthologs for *SoxB1* and *SoxB2* confidently in cnidarians and protostomes. To improve the phylogenetic resolution, we followed two strategies. First, to augment the phylogenetic signal, we increased the number of positions included in the alignment by aligning only SoxB full sequences (plus some SoxE proteins as outgroups). This way, we could generate confident alignments not only of the short High Movility Group (HGM) domain but also of the C terminus. Second, to avoid long-branch attraction (LBA) artifacts, we excluded the *SoxB* from insects [in which these genes appear to be more divergent, probably because of the significant expansion of the SoxB family within this clade (4)] and those from other typically fast-evolving species (as tunicates and nematodes). We thus restricted our analysis to sequences from cnidarians, deuterostomes, and lophotrochozoan protostomes.

For species in which SoxB genes were not described previously or for which gene predictions at their respective genome browsers were fragmentary or poorly annotated, we performed tBLASTN searches online at the JGI webpage (http://genome.jgi-psf.org/euk_home.html) for *Branchiostoma floridae* JGI v1.0, *Capitella teleta* JGI v1.0, *Lottia gigantea* JGI v1.0, and *Xenopus tropicalis* JGI v4.1. We performed searches at the NCBI webpage (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi) for *Strongylocentrotus purpuratus* Build 2.1 and at the Baylor College of Medicine webpage (http://www.hgsc.bcm.tmc.edu/project-species-x-organisms.hgsc) for *Saccoglossus kowalevskii*. We then downloaded each corresponding genomic region and built different gene models using GenScan (5) and GeneWise2 (6) software if necessary. We compared these predictions with expressed sequence tags and existent gene models when available. In general, both SoxB1 and SoxB2 genes were intronless, rendering reliable predicted coding sequences. Accession numbers for the previously published sequences are as follows: *Acropora millepora*: *SoxB1* EU784831, *SoxBa* EU784832, *SoxBb* EU784833, *SoxE1* EU784835; *Danio rerio*: *Sox14* NP_001032769.1, *Sox21a* NP_571361.1, *Sox21b* NP_001009888.1; *Gallus gallus*: *Sox1* NP_989664.1, *Sox2* NP_990519.1, *Sox3* NP_989526.1, *Sox14* NP_990092.1, *Sox21* Q9W7R5; *Homo sapiens*: *SOX1* NP_005977.2, *SOX2* NP_003097.1, *SOX3* NP_005625.2, *SOX14* NP_004180.1, *SOX21* NP_009015.1, *SOX8* NP_055402.2, *SOX9* NP_000337.1, *SOX10* NP_008872.1; *Nematostella vectensis*: *SoxB1* DQ173695, *SoxB2* (previously published as *Sox1*) DQ173692, *SoxBa* (previously published as *SoxB2*) DQ173696, *Sox2* DQ173693, *Sox3* DQ173694, *SoxE.1* DQ173697; and *Platynereis dumerilii*: *SoxB2* (previously published as *SoxB*) CAY12631.

Amino acid sequences for the HMG-box and the C terminus of the genes were aligned using MAFFT (7, 8) as implemented in Jalview 2.4 (9). A phylogenetic tree then was generated by the Bayesian method with MrBayes 3.1.2 (10, 11), with the WAG model. Two independent runs were performed, each with four chains. Following convention, convergence was reached when the value for the SD of split frequencies remained below 0.01. Burn-in was determined by plotting parameters across all runs: All trees before stationarity and convergence were discarded, and consensus trees were calculated for the remaining trees (from at least 1,000,000 generations).

Finally, as a complementary approach to phylogenetic analysis, we performed global BLASTP searches using only the C terminus of all of the cnidarian *SoxB*-related genes (i.e., excluding the HMG-box) against the "Non-redundant protein sequences" database at the NCBI webpage (http://blast.ncbi.nlm.nih.gov/Blast.cgi). The resultant hits were scrutinized carefully for the types of *SoxB* retrieved in the analyses (Fig. S6).

Generation of Sea Urchin Enhancer Reporter Constructs. To test putative transcriptional enhancers in sea urchin, a reported vector we generated a reported vector using the *zgata2* promoter and the EGFP reporter gene. The efficiency and absence of ectopic expression of these constructs was analyzed by injecting the empty vectors (Fig. S7).

Cloning of Deeply Conserved CNRs. To clone the putative regulatory elements from the different species, primers were designed to span the whole conserved sequence plus ~100 nt on either side. All primer sequences are listed in Table S2. PCRs were performed on human, amphioxus (*Branchiostoma floridae*), *Saccoglossus kowalevskii*, sea urchin (*Strongylocentrotus purpuratus*), or *Nematostella vectensis* genomic DNA using Expand High-Fidelity System (Roche). Amplicons were cloned in pCR8GW/TOPO vector (Invitrogen) according to the manufacturer's recommended protocol. Verified clones then were transferred to the different destination vectors with the Gateway recombination system (Invitrogen). The ZED vector (12) was used for the zebrafish transgenic experiments, and the vectors described above were used for the transgenic assays in sea urchin. The final transgenic constructs were purified using phenol-chloroform and

were normalized at 50 ng/mL in diethylpyrocarbonate-treated water before microinjection.

Zebrafish Trangenesis and Morpholino Microinjections, Culturing, and Whole-Mount in Situ Hybridization. For zebrafish transgenesis, 50–100 pg of each ZED-CNR vector was injected into one-cell–stage embryos together with 50–100 pg of Tol2 mRNA. Injected fish were observed at 24 and 48 h postfertilization (hpf) for GFP expression driven by the enhancer activity. As internal injection quality control, muscle RFP expression was determined 72 hpf. Embryos then were selected and raised to sexual maturity. Three or more independent stable transgenic lines were generated for each construct. Zebrafish *sox21b* and *id1* were amplified using the oligos listed in Table S2. PCR fragments were cloned in pCR8/TOPO, and RNA probes were generated using T7 polymerase and standard procedures. Embryos were raised at 28 °C in standard E3 medium and fixed at different stages in 4% paraformaldehyde overnight at 4 °C. In situ hybridizations were carried out as described (13).

To knockdown *zsox21a*, we used the *MOzsox21a* (5′-CATTTCTTGATACTTTGGTGCTCCT-3′) previously reported (12). To knock down *zsox21b*, we used the *MOzsox21b* (5′-CGCGTTCACCTGCTTCAGGTAGAAA-3′). These morpholinos (MOs) were injected in one-cell–stage embryos at 15 ng (when injected individually) or at 7.5 ng (when injected together). To evaluate the specificity of *MOzsox21b*, we fused the 5′ UTR plus the first four amino acids of *zsox21b* with the entire GFP ORF. GFP protein is not observed in embryos coinjected with 500 pg of this *zsox21b*::GFP mRNA and the *MOzsox21b*, although it is clearly detectable in embryos injected with the mRNA only.

Microinjection of Sea Urchin Eggs, Culturing, and Whole-Mount in Situ Hybridization. CNR were microinjected into fertilized *Strongylocentrotus purpuratus* eggs as described (14). *S. purpuratus* larvae from injected and uninjected eggs were grown at 16 °C, fed on *Rhodomonas lens* and *R. salina*, and cultured as described (15). Larvae from injected eggs were collected at different stages for observation by fluorescence microscopy for qualitative assessment of spatial activity of GFP. Larvae from uninjected eggs were fixed at different stages in a mixture containing both 2% paraformaldehyde and 2% formaldehyde as described in ref. 16 for whole-mount FISH. Riboprobes were labeled with digoxigenin, anti–digoxigenin-peroxidase was used for immunoreaction, and fluorescence was coupled to the bound antibody with fluorescein tyramide.

***Drosophila* Experiments.** The *Drosophila* integration vector pBPUwdGFP was generated, using standard molecular biology techniques, by replacing the *Gal4* coding sequence from pBPGUw (17) with d2EGFP, a destabilized variant of GFP (pd2EGFP-1; Clontech). Transfer of Gateway clones (*HsSox21*CNR-pCR8/TOPO, *SpSoxB2*CNR-pCR8-TOPO, and *HsId*CNR-pCR8-TOPO) into the integration vector was performed according to the manufacturer's recommendations (Invitrogen). DNA constructs were microinjected into embryos from flies carrying the landing platform ZH- attP- 22A (18), using standard *Drosophila* transformation techniques. Single males derived from these embryos were crossed to *yw* females. Transformant males (*w*+) were outcrossed twice to *w; Sp*/*CyO* females, and balanced stocks were established. *Drosophila* embryo collections were fixed and stained according to standard procedures. Anti-GFP (Molecular Probes) was used to detect reporter gene expression. For immunohistochemical detection, a biotin-labeled secondary antibody, followed by the Vectastain ABC kit (VectorLlabs) was used to develop signal. Third-instar larvae brains were fixed for 20 min in 3.7% formaldehyde/PBS. Primary antibodies used were mouse anti-Dachshund and mouse anti-Prospero (Developmental Studies Hybridoma Bank), mouse DE-cadherin (BD Transduction Laboratories), and rabbit anti-GFP (Molecular Probes). Appropriate Alexa Fluor-conjugated secondary antibodies (Molecular Probes) were used. Standard protocols were used for fluorescent immunostaining, except that PBS/0.3% Triton X-100 was used in all steps. Imaging was performed with a Leica SP2 confocal setup.

Prediction of Transcription Factor Binding Sites in *SoxB2*-CNRs. Sequences corresponding to the injected *SoxB2*-CNR (plus zebrafish) were scanned using Match (BIOBASE Biological Databases GmbH, http://www.gene-regulation.com/pub/programs.html#match) and Consite (http://asp.ii.uib.no:8090/cgi-bin/CONSITE/consite). For Match, 0,99 and 0,85 were used as cutoffs for core and matrix similarity, respectively, and "vertebrates" was selected as group of matrices. For Consite parameters were established as >0.55 for conservation and >0.80 for transcription binding site scores. For the graphical representation, overlapping binding sites of different members of the same gene families were merged. Only predicted sites conserved in at least two species were selected; sites corresponding to taxonomically restricted gene families were discarded. Sequences also were scanned for Sox2 binding sites using the corresponding matrix model (ID: MA0143.1) in JASPAR (http://jaspar.genereg.net/), with 75% as the relative profile score threshold.

1. Putnam NH, et al. (2008) The amphioxus genome and the evolution of the chordate karyotype. *Nature* 453:1064–1071.
2. Magie CR, Pang K, Martindale MQ (2005) Genomic inventory and expression of Sox and Fox genes in the cnidarian *Nematostella vectensis*. *Dev Genes Evol* 215: 618–630.
3. Shinzato C, et al. (2008) Sox genes in the coral *Acropora millepora*: Divergent expression patterns reflect differences in developmental mechanisms within the Anthozoa. *BMC Evol Biol*, 10.1186/1471-2148-8-311.
4. Wilson MJ, Dearden PK (2008) Evolution of the insect Sox genes. *BMC Evol Biol*, 10.1186/1471-2148-8-120.
5. Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. *J Mol Biol* 268:78–94.
6. Birney E, Durbin R (2000) Using GeneWise in the *Drosophila* annotation experiment. *Genome Res* 10:547–548.
7. Katoh K, Kuma K, Toh H, Miyata T (2005) MAFFT version 5: Improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res* 33:511–518.
8. Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 30: 3059–3066.
9. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ (2009) Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25: 1189–1191.
10. Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17:754–755.
11. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–1574.
12. Bessa J, et al. (2009) Zebrafish enhancer detection (ZED) vector: A new tool to facilitate transgenesis and the functional analysis of cis-regulatory regions in zebrafish. *Dev Dyn* 238:2409–2417.
13. Tena JJ, et al. (2007) Odd-skipped genes encode repressors that control kidney development. *Dev Biol* 301:518–531.
14. Oliveri P, Tu Q, Davidson EH (2008) Global regulatory logic for specification of an embryonic cell lineage. *Proc Natl Acad Sci USA* 105:5955–5962.
15. Leahy PS (1986) Laboratory culture of *Strongylocentrotus purpuratus* adults, embryos, and larvae. *Methods Cell Biol* 27:1–13.
16. Arenas-Mena C, Cameron AR, Davidson EH (2000) Spatial expression of Hox cluster genes in the ontogeny of a sea urchin. *Development* 127:4631–4643.
17. Pfeiffer BD, et al. (2008) Tools for neuroanatomy and neurogenetics in Drosophila. *Proc Natl Acad Sci USA* 105:9715–9720.
18. Bischof J, Maeda RK, Hediger M, Karch F, Basler K (2007) An optimized transgenesis system for Drosophila using germ-line-specific phiC31 integrases. *Proc Natl Acad Sci USA* 104:3312–3317.

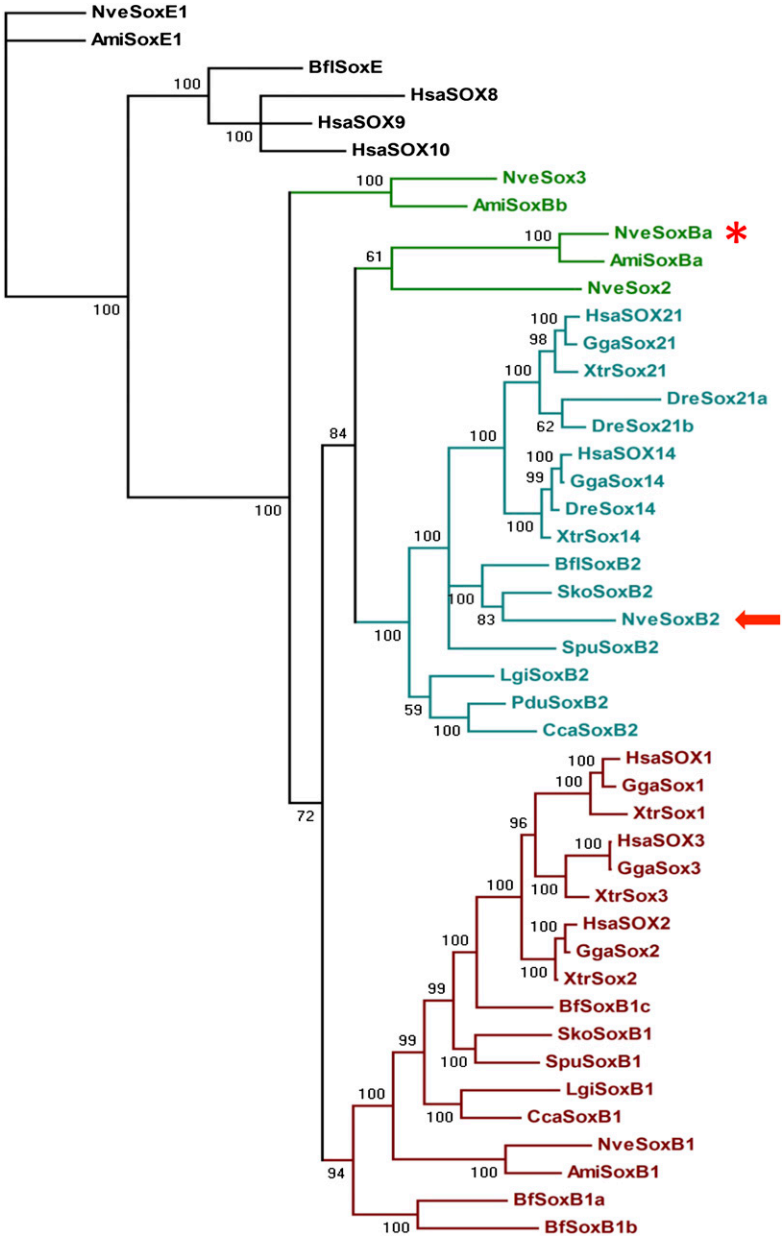


Fig. S1. Orthology relationships of *Nematostella SoxB* genes. Bayesian inference phylogenetic reconstruction of the *SoxB* genes from various eumetazoans. Eumetazoan *SoxB* genes are clustered into two different groups: the SoxB1 group (94% of posterior probability, 100% excluding the very divergent SoxB1a and SoxB1b from amphioxus, in red) and the SoxB2 group (100% of posterior probability, in blue). Therefore, the two known *SoxB* subgroups are present in deuterostomes, lophotrochozoan protostomes, and cnidarians, indicating that both subgroups have split in an ancient common ancestor of all eumetazoans. Furthermore, this analysis shows that the gene linked to the *SoxB2*-CNR in *Nematostella*—previously known as "*NvSox1*" (1) and renamed here "*NveSoxB2*" (red arrow) is the clear ortholog of the bilaterian *SoxB2* genes. In addition, a global BLAST search was performed using only the C terminus of the *Nematostella* SoxB2 protein (i.e., excluding the HMG domain), and it specifically retrieved SoxB2 from different bilaterians with highly significant e-values (up to 6×10^{16}). As in previous analyses (1, 2), the phylogenetic relationships among other *SoxB*-like genes from *Nematostella* and the coral *Acropora millepora* (in green font) could not resolved with enough confidence in the present analysis. Global BLAST searches using the C terminus of the proteins also were not informative, with none of these cnidarian *SoxB*-like genes retrieving any single Sox protein. One of the genes in this group (red asterisk) had been postulated previously to be the putative *Nematostella* ortholog of *SoxB2* (1). To avoid confusion, we rename it here as "*NveSoxBa*," following the nomenclature adopted in ref. 2 for its ortholog in *A. millepora*. Species are abbreviated as follows: Ami, *Acropora millepora*; Bfl, *Branchiostoma floridae*; Cte, *Capitella teleta*; Dre, *Danio rerio*; Gga, *Gallus gallus*; Hsa, *Homo sapiens*; Lgi, *Lottia gigantea*; Nve, *Nematostella vectensis*; Sko, *Saccoglossus kowalevskii*; Spu, *Strongylocentrotus purpuratus*; Xtr, *Xenopus tropicalis*. SoxE genes were used as outgroups.

1. Magie CR, Pang K, Martindale MQ (2005) Genomic inventory and expression of Sox and Fox genes in the cnidarian Nematostella vectensis. *Dev Genes Evol* 215:618–630.
2. Shinzato C, et al. (2008) Sox genes in the coral Acropora millepora: divergent expression patterns reflect differences in developmental mechanisms within the Anthozoa. *BMC Evol Biol* 8:311.

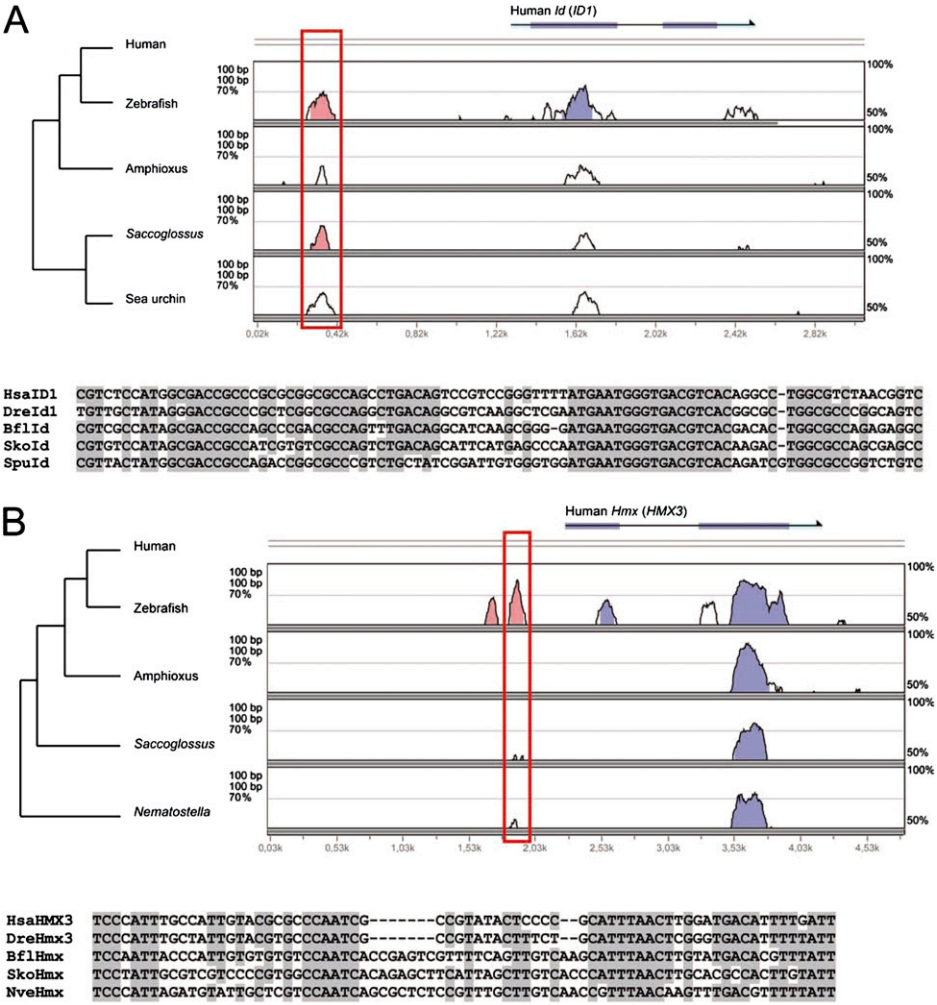
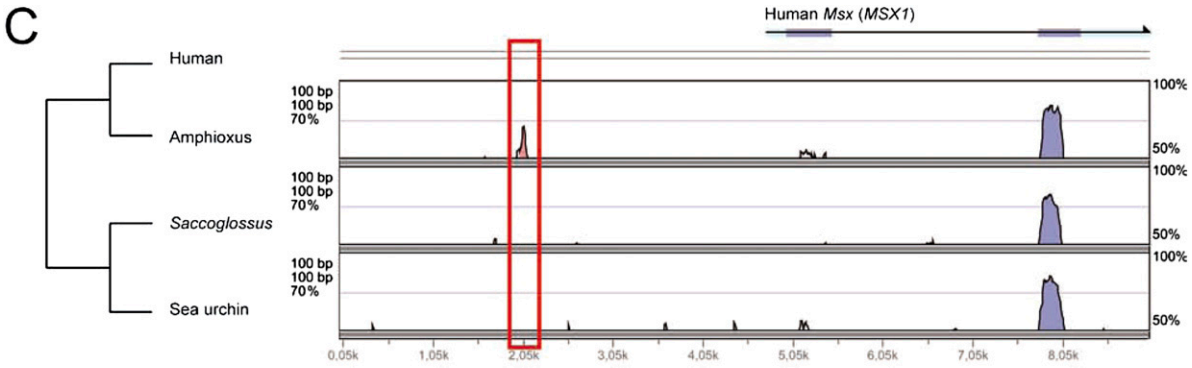
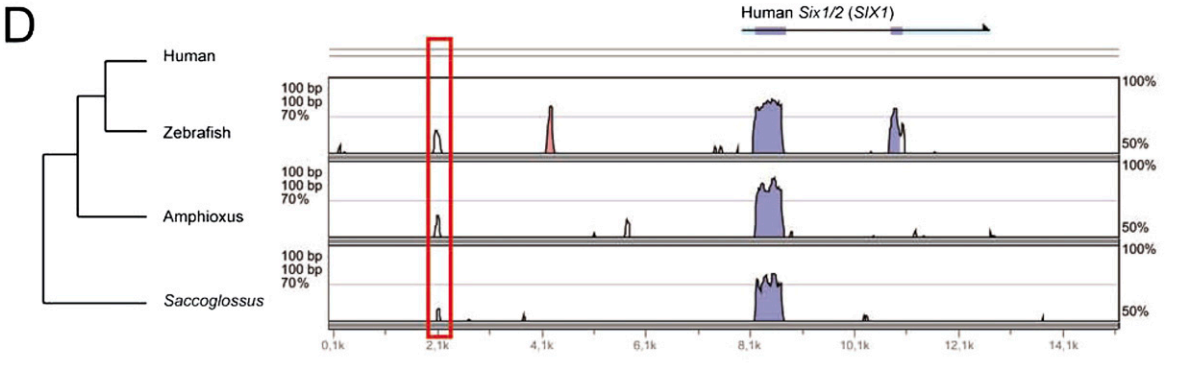


Fig. S2. (Continued)




```
HsaMSX1  CTTTGATCCGGCCGCCG----------CTGGCGCCTTAAAACAACATC---CTAGTCTGCCTATT-----
BflMsx   CTTTGATCCCGCCGTTC----------GTGGCGCCTAGAAACAACACAA--CTTGTCTGCGCATT-----
SkoMsx   CTTTGATCGGGCCGTTAACGCTCTCTTAGTGGCGCCTGTACACTCACTAAAGGATGTCTGCTTTTTTTTAA
SpuMsx   CTTTGATTGGGGGAGAG--------TT-GTGGCGCCGACTGCCGTCCTAA--GATGTCTACTCAGT-----

HsaMSX1  AGGCGCGCTGAGGAATCGTGACACATTGTTTATCCCCGCAATTAGC
BflMsx   GAAATATCTGACAAATCAGGTCAGATTGTTTATCCTCACAATTAGT
SkoMsx   AAAAAGGATGGCAAGTCGTGATGGATTATTTGTCGAGCCAATTAGC
SpuMsx   GAATAAGGTGACAAATGGTGACGGAGTGTTTGTACTTGCAATTAGC
```




```
HsaSIX1   TTGTCCCCCTGGCATCTCAGGTTAAAGAAGGAGGTCCAGGAAAATGTGGTAGAGCCATTACAAGAAACCCGAGCCGCAGTCA-TCA
DreSix1b  ACGGCCTTTCGATTTCTGAGGTTTAAGGAGGGGGCGCGCCGCAATGTGGTGGAGACATTACGA-CAACCCGAGCGCC-GTCAATCA
BflSix1/2 TCGGCCCTCTGCCGGCTCGGGTTGTAGGGGGAGGTTTAG-AAAATATGGTATAGTCATTAAGCGAAATTCGATACCCATTCA-TCA
SkoSix1/2 TCAGCCCTCTCGCGTCTCGGGTTGAAGTAGGAGGTTTAA-GAAATATGGCATAGTCATTAAGGGTAATTCGATACCCATTCA-CGA
```

Fig. S2. (Continued)

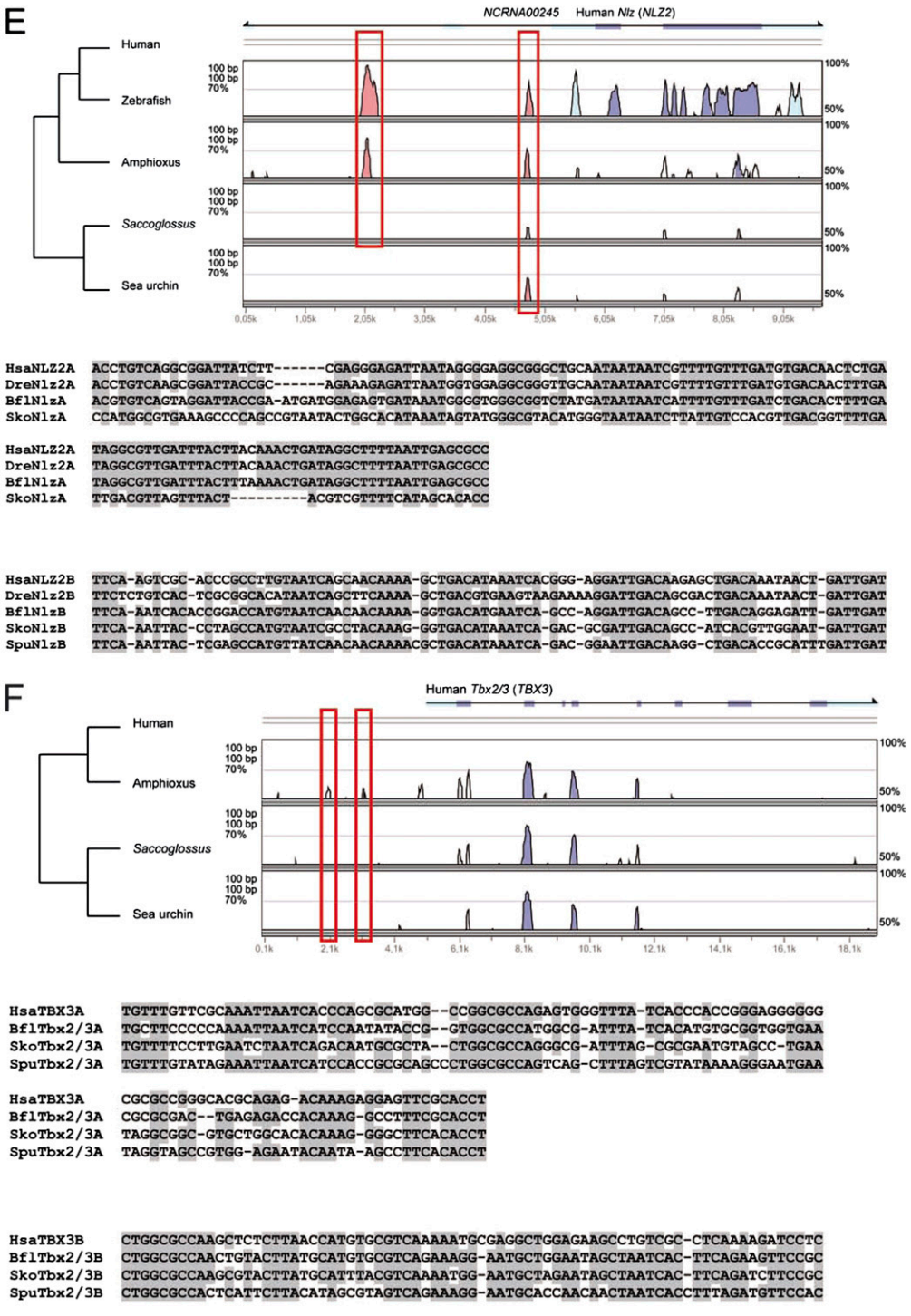


Fig. S2. Deeply conserved noncoding regions associated with developmental genes. Upper panels show VISTA alignment of *Id1* (*A*), *Hmx* (*B*), *Msx* (*C*), *Six1/2* (*D*), *Nlz* (*E*), and *Tbx2/3* (*F*) loci in different species. Blue/pink peaks in VISTA alignments correspond to coding/noncoding regions, respectively. Red rectangle indicates the CNRs. The lower panel shows sequence alignment of the different CNRs. Bfl, amphioxus; Dre zebrafish; Hsa human; Nve, *Nematostella*; Sko, *Saccoglossus*; Spu, sea urchin.

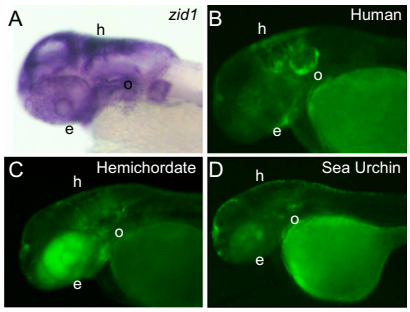


Fig. S3. *Id*-CNRs transgenic zebrafish promote expression at endogenous gene-expression domains. Images show lateral views of zebrafish embryos at 48 hpf. (*A*) Endogenous zebrafish *id1* expression pattern. (*B–D*) *Id1*-CNRs show largely overlapping expression domains in the eye (e), dorsal hindbrain (h), and otic vesicle (o) that overlap with endogenous *zid1* expression.

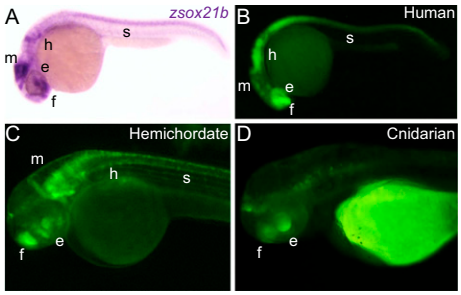


Fig. S4. *SoxB2*-CNRs promote expression at endogenous gene-expression domains in transgenic zebrafish. Images show lateral views of zebrafish embryos at 30 hpf (*A* and *B*) and 48 hpf (*C* and *D*). (*A*) z*sox21b* is expressed in the forebrain (f), eye (e), midbrain (m), hindbrain (h) and spinal cord (s). (*B*) *HsSOX21*-CNR promotes strong GFP expression in all these tissues except the midbrain, a region that expresses only low levels of this reporter gene. (*C* and *D*) *Saccoglossus* (*C*) and *Nematostella* (*D*) *SoxB2*-CNRs show similar but not identical expression domains. Note the weaker expression observed for the *Nematostella SoxB2*-CNR in most tissues except the forebrain and the eye.

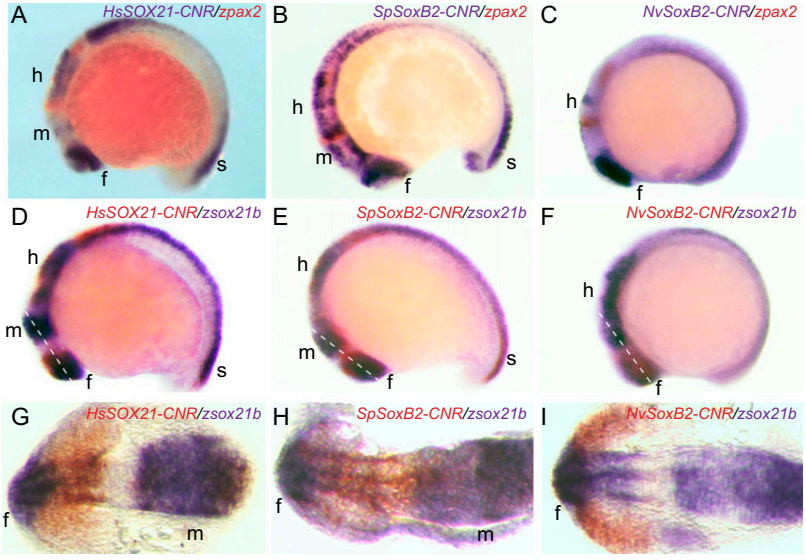


Fig. S5. Expression promoted by orthologous *SoxB2*-CNRs at the 5–15 somites stage. (*A–F*) Lateral views of embryos at the 5–15 somites stage double stained for the combinations of genes and GFP promoted by the distinct *SoxB2*-CNRs shown in the upper right corner of each panel. (*G–I*) Sections from the double stained embryos shown in panels *A–F*. The plane of these sections is shown by dotted lines in *D*, *E*, and *F*, respectively. f, forebrain; h, hindbrain; m, midbrain; s, spinal cord.

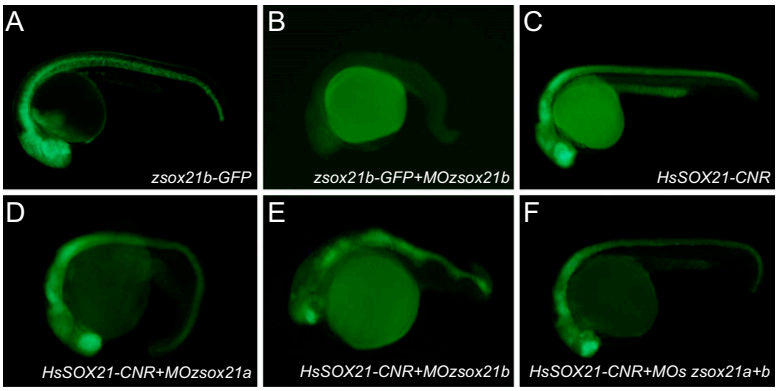


Fig. S6. *SoxB2*-CNR activity in not affected in *zsox21*-depleted embryos. All panels show GFP expression in lateral views of 24-hpf embryos. (*A* and *B*) GFP was strongly expressed in embryos injected with 500 pg of the *zsox21b-GFP* mRNA (*A*), but this expression was not detectable when this mRNA was coinjected with 15 ng of *MOzsox21b* (*B*). (*C*) GFP expression in *HsSOX21*-CNR transgenic embryos. (*D–F*) GFP expression is not altered in embryos injected with *MOzsox21a* (*D*), *MOzsox21b* (*E*), *or MOzsox21a+ MOzsox21b* (*F*).

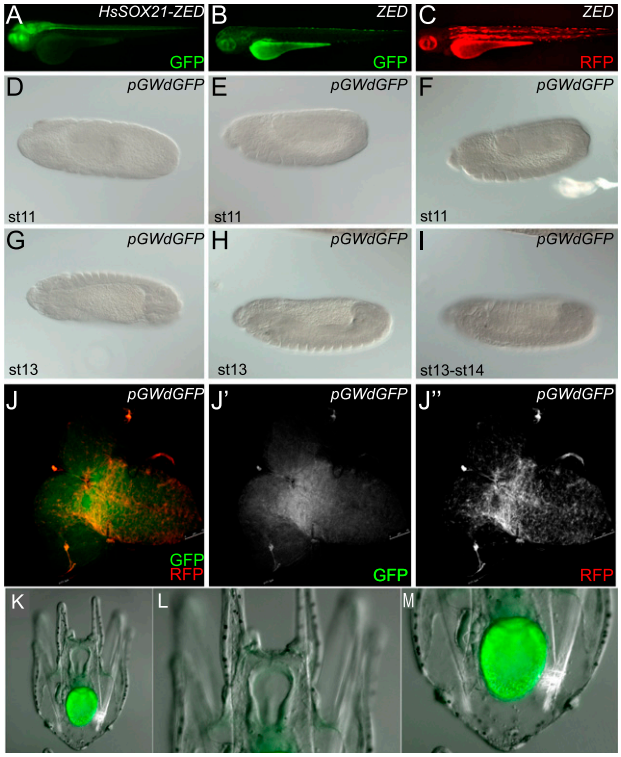


Fig. S7. Representative control transgenic embryos harboring empty reporter vectors. (*A–C*) Lateral views of 72-hpf zebrafish transgenic embryos. (*A*) GFP expression in brain promoted by human *SOX21-CNR*, as in examples in text. (*B* and *C*) GFP (*B*) and RFP (*C*) expression in a control stable transgenic embryo generated with the empty ZED vector. No GFP expression above background level could be observed in brain (fluorescent ovoid body is autofluorescence from contents of stomach). RFP expression is promoted by the cardiac actin promoter and is used as a positive control of transgenesis. (*D–I*) Control *Drosophila* transgenic lines containing the empty transformation vector pBPUwdGFP. No expression of GFP, driven by promoter only, could be detected in embryos after immunostaining against GFP. Embryos are shown in dorsal (*D* and *G*), lateral (*E* and *H*), and dorsolateral (*F* and *I*) views. (*J–J''*) In third-instar larval brains there is no cell-specific GFP expression (images were overexposed), but we do detect the RFP expression that is associated with the transgenesis landing site. (*K–M*), GFP expression driven by empty vector in sea urchin larvae. (*K*) Low-magnification view of whole larva, in which only autofluorescence of algae in stomach is evident. (*L* and *M*) High-magnification views of innervated rim of oral field where the same vector driven by the soxB CNR expresses in neurons, as discussed in text. As expected, no such activity is produced by the empty vector. Green staining observed in *K* and *M* correspond to endogenous autofluorescence.

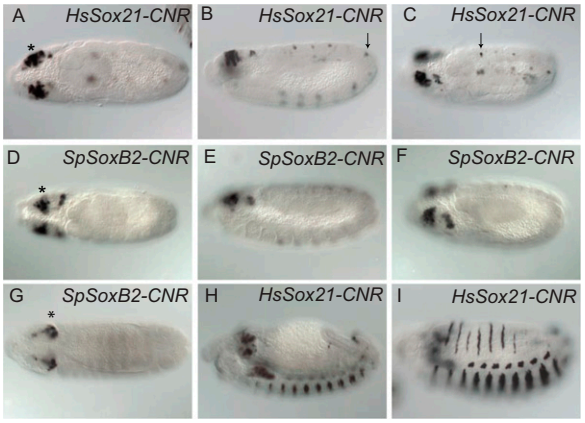


Fig. S8. Enhancer activity of *HsSox21* and *SpSoxB2* CNRs in *Drosophila* embryonic neuroectodermal derivatives. (*A–F*) Expression in *Drosophila* transgenic embryos of destabilized GFP (dGFP) driven by *HsSox21*-CNR (*A–C*) and *SpSoxB2*-CNR (*D–F*), detected with an anti-GFP antibody. *SoxB2*-CNRs drive expression of dGFP in neuroblasts of the presumptive brain (asterisks) and in ventral nerve cord neuroblasts (arrows). Dorsal (*A* and *D*), lateral (*B* and *E*), and dorsolateral (*C* and *F*) views of stage 11 embryos are shown. (*G–I*) At later developmental stages *SoxB2*-CNRs activity is detected in the embryonic brain (asterisk, stage 13 embryo) (*G*), ventral and dorsal epidermal stripes (*H* and *I*), and chordotonal organs (stage 15 embryo) (*I*).

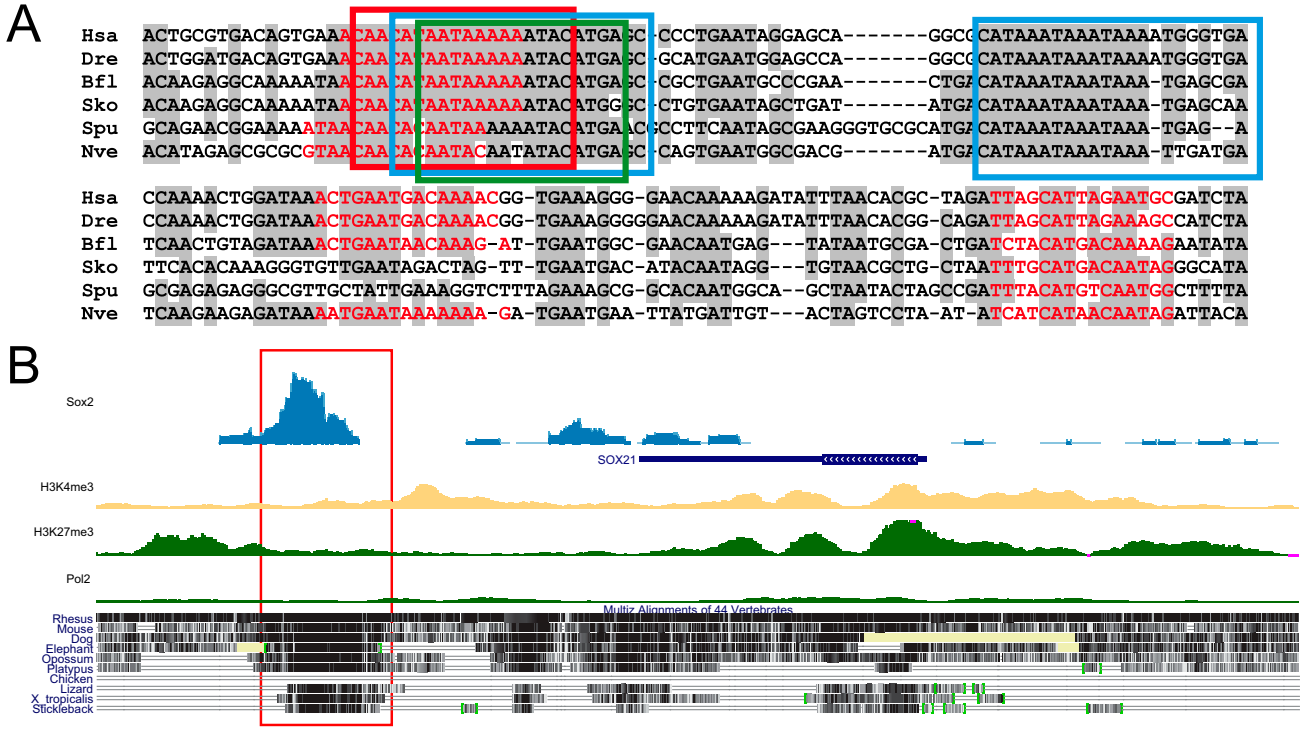


Fig. S9. Potential upstream factor acting in *SoxB2*-CNRs. (*A*) Transcription factor binding sites at *SoxB2*-CNRs, as predicted by the Consite and Match programs. Red, blue, and green boxes mark predicted binding sites for the Sox (including D, E, and F groups), Fox (including A, D, F, I, L, and Q families), and Pou3f family of transcription factors, respectively. Nucleotides in red indicate putative binding sites for Sox2 that are conserved between at least two species, as predicted by the Sox2 Matrix model in JASPAR. Bfl, amphioxus; Dre, zebrafish; Hsa, human; Nve, *Nematostella*; Sko, *Saccoglossus*; Spu, sea urchin. (*B*) Distribution of Sox2-bound sites at the *SOX21* locus in human stem cells. The Sox2 ChIP-sequencing (ChIP-Seq) signal is strong at the *SOX21*-CNR (red box). In these cells the *SOX21* promoter shows bivalent H3K4me3 and H3K27me3 epigenetic marks, indicative of a transcriptionally poised state. Accordingly, in these cells very little transcription of *SOX21* gene is detected, as determined by Pol2 ChIP-Seq. H3K4me3, H3K27me3, and Pol2 ChIP-Seq profiles were obtained from the ENCODE data at the University of California, Santa Cruz browser (1). The Sox2 genome-wide ChIP-Seq profile was obtained from ref. 2.

1. Kent WJ, et al. (2002) The human genome browser at UCSC. *Genome Res* 12:996–1006.
2. Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32:1792–1797.

Table S1. Studied chordate CNRs and their phylogenetic distribution across major metazoan groups

CNR	Gene	Gga*	Xtr*	Dre*	Tru*	Bfl	Sko	Spu	Dpu	Cte	Lgi	Nve	Tad	Human location	Amphioxus location
1	*SOX21*	N.A.	YES	N.A.	YES	YES	YES	YES	NO	NO	NO	YES	NO	chr13:94156903–94157110	scaffold_50:1169971–1170173
2	*MSX1*	Yes	Yes	No	Yes	Yes	Yes	Yes	No	No	No	No	N.D	chr4:4976727–4976823	scaffold_56:2362024–2362121
3	*SIX1-a*	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No	No	chr14:60191711–60191798	scaffold_52:2878536–2878622
4	*SIX1-b*	No	No	Yes	Yes	Yes	No	No	No	No	No	No	No	chr14:60188067–60188154	scaffold_52:2880653–2880738
5	*ID1*	Yes	N.A.	N.A.	Yes	Yes	Yes	Yes	No	No	No	No	No	chr20:29655769–29655845	scaffold_166:1326698–1326773
6	*ZIC5*	N.A.	Yes	No	No	Yes	No	No	No	No	No	No	No	chr13:99413906–99413998	scaffold_40:1809962–1810050
7	*HMX3*	No	No	Yes	Yes	Yes	Yes	No	No	No	No	Yes	No	chr10:124885142–124885219	scaffold_406:567486–567572
8	*TBX3*	Yes	Yes	N.A.	N.A.	Yes	Yes	Yes	No	No	No	No	No	chr12:113586552–113586623	scaffold_89:309180–309250
		Yes	Yes	N.A.	N.A.	Yes	Yes	Yes	No	No	No	No	No	chr12:113587672–113587734	scaffold_89:310103–310164
9	*ONECUT1*	N.A.	Yes	No	No	Yes	No	No	No	No	No	No	N.D	chr15:50877303–50877377	scaffold_66:258311–258385
10	*NKX6.1*	N.A.	Yes	No	No	Yes	No	No	No	No	No	No	No	chr4:85757666–85757728	scaffold_294:852642–852707
11	*NLZ2-a*	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No	No	N.D	chr10:76835035–76835166	scaffold_9:3320557–3320693
	(NLZ1)	N.A.	Yes	Yes	No									chr8:37652068–37652163	scaffold_9:3320560–3320656
12	*NLZ2-b*	No	Yes	Yes	Yes	Yes	Yes	Yes	No	No	No	No	N.D	chr10:76832361–76832456	scaffold_9:3319055–3319149
13	*SIX1-c*	Yes	Yes	No	No	Yes	No	No	No	No	No	No	No	chr14:60180747–60180808	scaffold_52:2889041–2889107

Gene corresponds to the gene to which the element is physically linked. Yes/No indicates presence/absence in the corresponding genome. N.A. indicates the sequence was not available in the automatic Vista alignments. N.D. No clear ortholog of the associated gene could be identified in the genome. Cte, *Capitella teleta*; Dpu, *Daphnia pulex*; Dre, *Danio rerio*; fl, *Branchiostoma floridae*; Gga, *Gallus gallus*; Lgi, *Lottia gigantea*; Nve, *Nematostella vectensis*; Sko, *Saccoglossus kowalevskii*; Spu, *Strongylocentrotus purpuratus*; Tad, Trichoplax adhaerens; Tru, *Takifugu rubripes*; Xtr, *Xenopus tropicalis*.
*From automatic VISTA alignments.

Table S2. Primers used to amplify CNRs and to generate probes for in situ hybridization

Sox21b-CNR	Forward (5′-3′)	Reverse (5′-3′)
Human	ACATCCCGAGGCAGCA	CCTTCAGTGGGAGCGTTTA
Amphioxus	CCGCCAGATGGATTCTAAAG	GGAAAGTCACACACGGAAATC
Saccoglossus	GCTTCCAGTTTTCCCTCGTTGCCCG	CCAATATGGCGAGACAGGGACACG
Sea urchin	ATCCCGAAACCTCTCCATCT	GGCCATGATAGGGCAAAACT
Nematostella	GTTTGCAGTCTTCTTTAGAAATGT	CCGTTTGGGAATAATTCTCTCATT
Id1-*CNR*		
Human	GTGAAGAAACCCCAAGCG	CCCATTTTTGGCTGCTT
Amphioxus	CCGGCCCATGAGTAACAGA	TGGGTCCCGATAGTCTGATTT
Saccoglossus	CGTCCAGTCTCGATGACCTA	AGGGGGCAGTCTACAAAACC
Sea urchin	CAAGCGCAGACCTAACACAA	ATACACCTCCCCCGAGACTT
Msx1-CNR		
Human	GCCTAGCAACGTTTACACG	GGGCGGCCTGAAAAG
Six1_a-CNR		
Human	TTATTCCATGCAATTAGATCTTG	CCTATTTGTGTCTAATTGCCATC
Six1_b-CNR		
Human	AATCAGGCTTCGTGAAATTTG	CTTCATCTCGGTCCACCAG
Hmx3-CNR		
Human	AGGTCACGTCCTGCCTCTT	GCTGGGGTCCCAGAACT
Tbx3-CNR		
Human	CTCTAGGCATCGCACTTATCT	CTCTATTTTAACCACTGGCAGAG
NZL2b_a-CNR		
Human	TTCGCTTTTCTCAAAACTCC	CCAGGCTTCTCCCCTG
NZL2b_a-CNR		
Human	AGCCACCAGGGTAAGAAGG	TCTCCCGTGTTAGAAGTTGC
Probes		
Zebrafish *Sox21b*	GGATCCAGCCACTATTTTTCCAGGATTTACC	CAGAGCTCTAACGCCGCCGCGTACG
Zebrafish *Id1*	GTCATCGCACTATCGACAACT	CACAATAAAGCGTTCACATCATAT