Nucleic Acids Research, 2005, Vol. 33, No. 17 ***5437–5445***
doi:10.1093/nar/gki853

Human-zebrafish non-coding conserved elements act *in vivo* to regulate transcription

Jordan T. Shin*, James R. Priest[1,2], Ivan Ovcharenko[3], Amy Ronco, Rachel K. Moore, C. Geoffrey Burns and Calum A. MacRae

Cardiovascular Research Center and Cardiology Division, Massachusetts General Hospital and Harvard Medical School, Charlestown, MA 02129, USA, [1]Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA, [2]DOE Joint Genome Institute, Walnut Creek, CA 94598, USA and [3]Lawrence Livermore National Laboratory, 7000 East Avenue, Livermore, CA 94550, USA

Received August 4, 2005; Accepted September 5, 2005

ABSTRACT

Whole genome comparisons of distantly related species effectively predict biologically important sequences—core genes and *cis*-acting regulatory elements (REs)—but require experimentation to verify biological activity. To examine the efficacy of comparative genomics in identification of active REs from anonymous, non-coding (NC) sequences, we generated a novel alignment of the human and draft zebrafish genomes, and contrasted this set to existing human and fugu datasets. We tested the transcriptional regulatory potential of candidate sequences using two *in vivo* assays. Strict selection of non-genic elements which are deeply conserved in vertebrate evolution identifies 1744 core vertebrate REs in human and two fish genomes. We tested 16 elements *in vivo* for *cis*-acting gene regulatory properties using zebrafish transient transgenesis and found that 10 (63%) strongly modulate tissue-specific expression of a green fluorescent protein reporter vector. We also report a novel quantitative enhancer assay with potential for increased throughput based on normalized luciferase activity *in vivo*. This complementary system identified 11 (69%; including 9 of 10 GFP-confirmed elements) with *cis*-acting function. Together, these data support the utility of comparative genomics of distantly related vertebrate species to identify REs and provide a scaleable, *in vivo* quantitative assay to define functional activity of candidate REs.

INTRODUCTION

Most vertebrate genomic sequence is neither transcribed nor encodes protein. However, 2- to 3-fold more sequence is estimated to be under selective pressure that can be accounted for by coding sequence alone (1,2). The ongoing generation and comparison of diverse vertebrate genomes reveals novel functional sequences based solely on their conservation throughout evolution (1,3). Defining important non-coding (NC) regions and predicting the biological function of such sequences will require sophisticated systems for empiric testing. *In vitro* approaches fail to represent the complete repertoire of transcriptional programs available *in vivo*. Ultimately, a systematic strategy combining *in silico* identification with biological validation in intact organisms will be required to annotate functional NC sequences.

Evolutionarily conserved regions (ECRs) are found in both coding and NC regions and have been identified by computational approaches using (length) windows of 70–100 bp and thresholds ranging from ~70% conservation (4–9) to complete identity (10). While coding ECRs are more easily defined by the presence of open reading frames (ORF), which can be tested for transcription, the function of NCECRs is less readily determined. NC conserved regions potentially may regulate many aspects of genome biology including RNA genes, enhancers, silencers or boundary elements.

Genome sequence alignments between human and rodent species (which diverged ~70 Mya) indicate extensive regions of similarity. Whole genome comparisons between human and mouse have found that 40% of these genomes can be aligned with each other, but only ~5% of these genomes are under

*To whom correspondence should be addressed. Tel: +1 617 724 9566; Fax: +1 617 726 5806; Email: jshin1@partners.org
Correspondence may also be addressed to Calum A. MacRae. Tel: +1 617 726 4343; Fax: +1 617 726 5806; Email: cmacrae@partners.org
Present address:
James R. Priest, Stanford University School of Medicine, Stanford, CA 94305, USA

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

© The Author 2005. Published by Oxford University Press. All rights reserved.

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article for non-commercial purposes provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated. For commercial re-use, please contact journals.permissions@oxfordjournals.org

active selection (1). These data suggest that there may be insufficient evolutionary divergence between humans and rodents to facilitate resolution of conserved, meaningful DNA from similar, yet irrelevant sequences. The specificity of regulatory sequence identification based on comparative genomic prediction is significantly increased when comparing the human genome to that of more distant species such as birds, amphibians and fish. To date, the most remote vertebrate genomes that have been compared are those of human and *Fugu rubripes* (Fugu) (2,3,11–13). However, the highly compact pufferfish genome (365 million bp) lacks a significant quantity of the intergenic NC sequence found in other species.

The zebrafish offers not only a divergent genome for comparative analysis, but also is an extremely tractable experimental system. Multiple strategies have been used to explore the potential of NC sequences to regulate gene expression *in-cis*. Transgenic mouse embryos expose the target sequence to a broad range of transcriptional programs but are resource intensive. Both stable and transiently transgenic zebrafish embryos have been employed to explore *cis*-regulation of gene expression on a gene–by–gene basis (14–24). The utility of the zebrafish positions the organism as a robust and simple system amenable to moderate and large-scale transcriptional regulatory analyses *in vivo*. Expression of a fluorescent reporter gene has been employed in zebrafish embryos to explore the spatio-temporal impact on expression during development of NCECRs from a human-pufferfish genomic comparison (13). We present a novel comparison between human and zebrafish draft genome sequences and extend the analysis of transcriptional regulation in the zebrafish to a quantitative, scaleable system using luciferase transgenesis.

MATERIALS AND METHODS

Human-Zebrafish genome alignments and dataset comparisons

Zebrafish sequence data (WTSI Zv3) were produced by the Zebrafish Sequencing Group at the Sanger Institute (http://www.sanger.ac.uk/Projects/D_rerio/). Human (NCBI build 35) and zebrafish genomes were aligned, and the alignments annotated and binned as described previously (25). We defined coding sequences as ECRs overlapping with the known genes, such as, human mRNA and xeno mRNA tracks at the UCSC browser (26). Clusters were defined when a sliding window of 150 kb contained a maximal number of NC conserved elements. Gene Ontology (GO) annotations and subsequent analysis were based on the annotation of the closest gene in the aligned human segment (27). To ensure a true and consistent comparison between organisms, a human-fugu alignment was performed using the same parameters and filtering criteria as the human-zebrafish dataset while the UC dataset was publicly available (http://www.sciencemag.org/cgi/content/full/1098119/DC1). Datasets were compared to each other with regard to the human anchor sequence. Conserved regions from different organisms were considered as overlapping if they are aligned to the same human sequence.

Cloning of human and zebrafish test fragments

A set of 16 conserved NC sequences were chosen for enhancer testing based on proximity within 100 kb (or less) of genes with known function(s). Primers were designed (Table 1) from repeat-masked genomic sequence to flank the element by 250–500 bp on either side (28). Negative control elements were chosen from human non-conserved, NC sequence. Sequences were amplified via PCR from commercial

Table 1. Experimental and positive control locations, DNA of origin, overlap with other dataset locations, right and left flanking primers

Name	Species	Location	Overlap to UC	L primer (5′→3′)	R primer (5′→3′)
HZ.H2	Human	chr15: 94,589,011-094.591,049	None	TCCAGGCAATAATGAGAAAGG	GAATGCTGGGAAAAGGAGAG
HZ.H3	Human	chr13: 76,963,964-976,965,840	None	TTGTGCCATCAGAGTCTTGC	AAATCCAGGCAGCTGACATT
HZ.H4	Human	chr16: 78,290,409-478,292,298	None	TACGATGGGATTGTGTCTGC	CAGCTTATTCAGAAAGGGCTTG
HZ.H5	human	chr4: 80,764,529-580,766,417	None	TCCTCCTTGTGTTCATTTCTTG	TTTTCACTTTTTCCCCCTTAC
HZ.H6	Human	chr10: 8,085,826-828,087,770	None	TTTTGTTCCTTCGGCGTTAG	GTGTCCAGATCTCCACGATG
HZ.H7	Human	chr4: 112,127,925-112,129,834	None	GTCACTCTTTGGGCTGGATG	CACGATGCTTTCAGAAATGTG
HZ.H8	Human	chr18: 51,238,041-051,240,065	uc.435+, uc.388+	GGTACCAGGTTGGCATCAAG	ACAGGGGGATTATGAAGACG
HZ.H9	Human	chr8: 106,567,420-106,569,441	None	GCTACCTCACTTCACGCTTTC	TTCCCTAATGCTTTTTAACTTGC
HZ.H10	Human	chr15: 65,739,790-765,742,119	None	CAGGCTGTGCATTCTACCTG	AAGCAAATGCCACCTACAGC
HZ.H11	Human	chr17: 35,465,051-035,467,078	None	TTGGCTAGGGGTAGCAGTTG	AATCCCAAGGGTCCCATAAC
HZ.H12	Human	chr1: 87,244,111-187,246,143	uc.29+	TCTGGCGTGTGACTATCTGG	TTATGGGCCCAGATTCAATG
HZ.Z1	Zebrafish	chr13: 7,685,256-7,687,093	None	CACCCAAGCACTGAGCGTATTCCA	GCCACACAATTGAAGCCTTT
HZ.Z2	Zebrafish	chr18: 27,546,874-27,549,311	None	CACCTCAATTGTTGCCGTAGTCCA	CCGTCATGCATTAGGTGTTG
HZ.Z3	Zebrafish	chr14: 29,552,127-29,553,616	None	CACCCGAGCTCGGTACCCTAATTG	ATTAATCGCGTTTGCTGAT
HZ.Z14	Zebrafish	chr25: 17,004,520-17,002,621	None	CACCAAACGAAGAACGGGGACTTT	GCGAGAGAAACGAAGGATTG
HZ.Z15	Zebrafish	chr19: 29,873,848-29,870,609	None	CACCCGCTCTGACCAAAGAGGTTC	TGTACACGCCAGGTTTAAGG
PC1	Human	chr13: 70,465,895-870,468,171	uc.351+	CGATTGCTTTCTCTTTTCCAG	GTCGGAAAGAGGCATCTCAG
PC2	Human	chr13: 70,233,780-770,226,388	None	GCACATGCCAAGTCCTGTC	GAACAAACTCTGGATTTTTGAGC
PC3	Human	chr13: 70,098,436-470,100,836	None	GCGGCCGCATTAGCAAAAGAAT-ACTTCCATGTCTGAG	ATGAAGAACCATCCCACTTG

The two letters to the left of the period in the name denotes that the sequences were selected from the HZ comparative dataset; the first letter to the right of the period denotes the genome of origin (H for human and Z for zebrafish). The absolute position in chromosomes (where that information was available) is listed in the location column. If there is overlap to the UC dataset (10) it is listed in the Overlap to UC column. Finally, specific primers used to amplify genomic DNA are listed in the last two columns. Zebrafish L-primer sequences all contain an initial CACC which was used for cloning.

preparations of zebrafish (Seegene Inc. Rockville, MD) or human placental (SigmaAldrich, St. Louis, MO) genomic DNA with Platinum *Taq* polymerase (Invitrogen Corp. Carlsbad, CA) and cloned directly into pENTR/D-Topo or Topo XL (Invitrogen Corp. Carlsbad, CA) before subcloning into the appropriate expression vectors.

Expression constructs

To assess expression we employed an 870 bp segment of the zebrafish cardiac myosin light chain 2 (cMLC2) promoter which is cardiomyocyte specific with essentially no extra-cardiac expression in either stable or transient transgenesis (24). Use of this promoter provides (i) an integrated, positive control for transgenesis and (ii) a negative background in non-cardiac tissues on which to visualize extra-cardiac transcription. For fluorescent assessment of *cis*-acting gene regulation, the Gateway Vector Conversion System (Invitrogen Corp. Carlsbad, CA) was used to introduce attR recombination sites into a GFP-containing reporter construct at a restriction site 5′ to cMLC2 promoter (29) to generate the plasmid pGM:GFP (Figure 1). For luciferase assays, the cMLC2 promoter was introduced into the multiple cloning site upstream of the coding sequences for firefly (FL) and renilla luciferase (RL) in pGL2 and pRL plasmids (Promega, Madison, WV). The attR recombination sites were also introduced to the 5′ promoter, to generate the plasmids pGM:GL2 and pGM:RL (Figure 1), respectively. Cloned conserved sequences were shuttled into the expression vectors either by recombination reaction (Invitrogen Corp. Carlsbad, CA) or blunt-ended ligation, and screened by PCR. Plasmid DNA was purified with QiaQuick columns (Qiagen, Valencia, CA).

Zebrafish care, husbandry and transgenesis

All zebrafish experiments were approved by the MGH Subcommittee on Research Animal Care. Zebrafish (AB strain) were raised and maintained using standard protocols. Purified plasmid DNA was suspended at 25 ng/µl in injection buffer [0.1% (w/v) phenol red (Sigma) in 0.3× Danieau buffer (17 mM NaCl, 2 mM KCl, 0.12 mM $MgSO_4$, 1.8 mM $Ca(NO_3)_2$ and 1.5 mM HEPES, pH 7.6)] for microinjection, immediately after fertilization (1–2 cell stage). Embryos were incubated at 28.5°C, and embryos which developed abnormally were removed after 24 h post fertilization (hpf).

Analysis of spatial expression

To examine spatial patterns of transcription, GFP reporter constructs containing each cloned conserved element were injected as above. A minimum of 30 transgenic embryos (those exhibiting cardiac GFP expression) were analyzed per construct. For monitoring of GFP expression, embryos were maintained with 0.003% phenylthiocarbamide to inhibit pigment formation, arrayed in 48-well plates and analyzed on an inverted microscope (TE200, Nikon) with fluorescent illumination and digital recording (ORCA-ER, Hamamatsu). Developing embryos were visually inspected for extra-cardiac GFP expression through 5 days of development and *cis*-regulatory activity was recorded; enumerating both the embryos with ectopic fluorescence and characterization of the tissues and cell type. The fraction of embryos exhibiting non-cardiac GFP expression was assessed for statistical significance using a χ^2-metric compared to pooled negative control results, applying the Bonferroni correction for multiple comparisons where necessary.

Analysis of luciferase expression

Expression constructs containing a cloned conserved sequence with the cMLC2 promoter and FL reporter cassette were resuspended in injection buffer with an equal amount of control vector containing the same promoter and RL to final concentration of 25 ng/µl. Embryos (15 per group) were harvested at 72hpf and mechanically disrupted in Passive Lysis Buffer (Promega). Lysates were cleared by centrifugation and supernatants analyzed serially for FL and RL according to manufacturers' protocols (Dual Luciferase Assay Kit, Promega) with a Victor 3 luminometer (Perkin-Elmer). Data are expressed as the mean ratio of a minimum of three samples ± SEM. Transcriptional regulatory activity of conserved sequences is compared to the pooled average luciferase ratio from constructs harboring negative control (human non-conserved NC) sequences ($n = 8$). Experimental expression ratios which exceeded the mean of the negative control ratios by at least 2 standard deviations were considered positive.

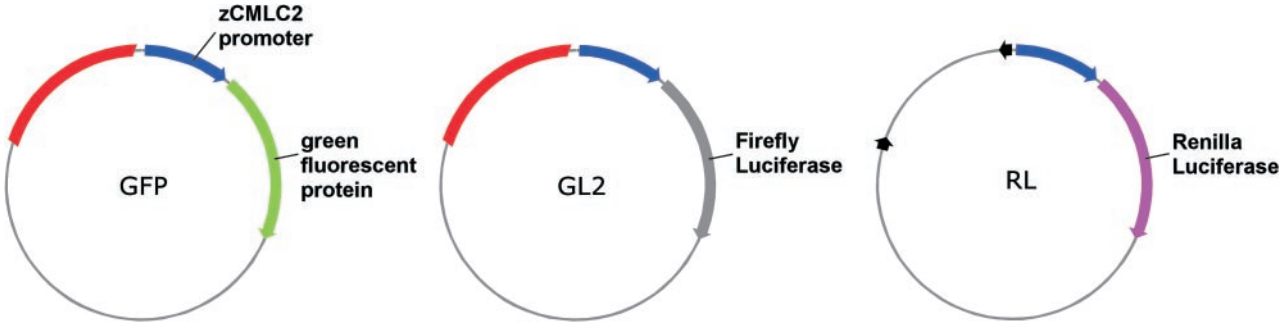

Figure 1. Expression vector constructs use for zebrafish transgenesis. The GFP construct is comprised of the zCMLC2 promoter (blue) driving expression of green fluorescent protein (green). Cloned sequences were inserted into the gateway recombination sequence (red) 5′ to the expression cassette and assessed for their ability to modulate transcriptional activity from this promoter. Similarly, the GL2 firefly luciferase construct consisted of an expression module containing the zCMLC2 promoter and the luciferase coding sequence (gray). The control pGM:RL plasmid, which contained the renilla luciferase coding sequence (purple), was used for normalization.

RESULTS

Human and zebrafish sequence comparison

Human:zebrafish (HZ) genome comparison using parameters of >70% identity and >80 bp in length identified a total of 6.5×10^4 conserved elements. We excluded sequences corresponding to known genes, as well as mRNA sequences from human and other species, in order to identify conserved NC sequences with potential gene regulatory activity. Filtering of the original HZ genomic comparison results in 4799 (7% of the original data set) conserved NC sequences (Figure 2A). A comparison of our dataset to recent reports describing 258 human ultra-conserved (UC) and human:pufferfish (HP) conserved elements (10,25), shows overlap between and among each of the data sets: 81 sequences are shared by all of the comparative data sets (31.6% of UC). Of the UC dataset 103 (40%) of the sequences are unique. Similarly, 40% (1197 of 2968 sequences) are unique in the HP set. The HZ comparison contains 3010 (63%) unique elements (Supplementary Table S1). Furthermore, the UC set retains substantial overlap with the HZ (49.2%) and HP (42.2%) elements (Figure 2B).

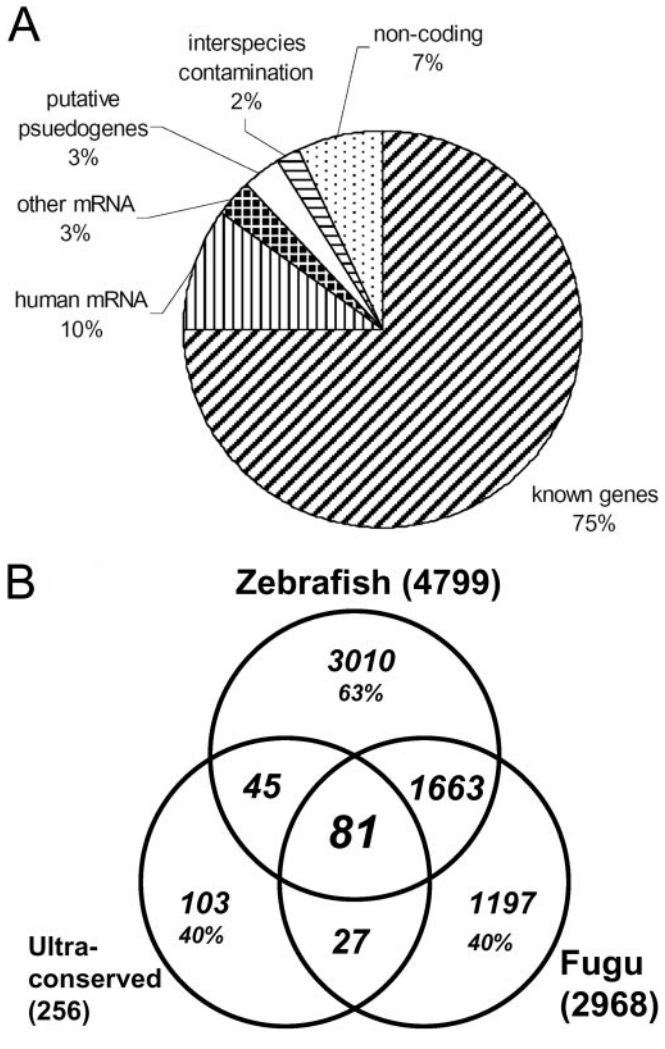

Figure 2. Bioinformatic characterization of HZ NC evolutionarily conserved regions and overlap with other genomic comparisons. (**A**) Breakdown and characterization of the 6.5×10^4 conserved regions shared between the human and zebrafish genomes. While most of the conserved sequences correspond to known genes and transcripts, 7% represent conserved, NC sequences. (**B**) A Venn diagram comparison of HZ, human:pufferfish and human:UC NC datasets. Where one species is indicated, the comparison is between human genomic sequence and that species. The UC set is that reported by Bejerano *et al*. (10). For each dataset, the number of total elements is indicated in parentheses while the percentage of total elements that are unique to that set is indicated in italics. The numbers within each segment of the Venn diagram represents the absolute number of elements.

While the expected average distance between a pair of HZ NCECRs is 500 kb, we detected 372 discrete clusters of 3 or more NCECRs in intervals smaller than 150 kb, comprising a full 79% of our dataset, consistent with previous observations (10,13,30). Interestingly, 126 clusters of NCECRs overlap with the gene deserts (12) in the human genome. Large clusters overlie the HSA19q gene desert and two others overlap *DACH1* gene and its associated gene deserts.

In order to address the possible biological function of the identified NCECR elements, we characterized the 290 genes that populate genomic regions neighboring clusters of NCECRs identified in our filtered set using GO annotation. Particular GO categories were found to be enriched in the list of genes nearest clusters of NCECRs. Specifically, over-represented terms include transcription factor activity (7.1 times in excess of the expected random frequency), regulation of transcription (3.9×), development (2.6×), morphogenesis (2.0×) and organogenesis (1.9×). In absolute terms, transcription factors represented the largest fraction (104 out of 290) of identified, neighboring genes.

Conserved NC sequences modulate organ-specific expression transient transgenesis

To test the biological potential of sequences identified by this comparative strategy, we employed a transient expression system in developing zebrafish embryos to detect *in-cis* modulation of organ-specific transcription by NCECRs. We generated an expression construct containing 870 bp of the zebrafish cardiac cMLC2 promoter, an ORF for GFP with an insertion site for cloned NC sequences 5′ to the promoter (Figure 2). A minimum of 100 embryos were injected with each construct and analyzed for transgene expression. Only transgenic embryos with detectable cardiac GFP expression were analyzed (average number 51; range 30–126).

Control transgenic embryos injected with either the promoter/reporter alone (pMLC:GFP) or with the Gateway sequence only demonstrated no extra-cardiac GFP expression (0%). Ectopic GFP expression from constructs with Gateway insertion sequences does not differ only from MLC:GFP constructs. Eight human non-conserved, NC sequences were cloned to serve as negative controls and inserted via recombination into the insertion site 5′ to the expression cassette. These constructs displayed minimal ectopic fluorescence (mean: 2.44 ± 0.74; range 0.00–05.56%) compared with the MLC:GFP only vector. Positive control sequences which had been shown to function as transcriptional regulators in transgenic mouse development were taken from previous comparative analyses (12). Two of three positive control enhancers functioned to augment transcription in our system, and displayed significant extra-cardiac GFP expression (Figure 3).

We randomly chose 14 NCECRs and in two cases also used both the fish and human paralogs (total 16 NCECRs).

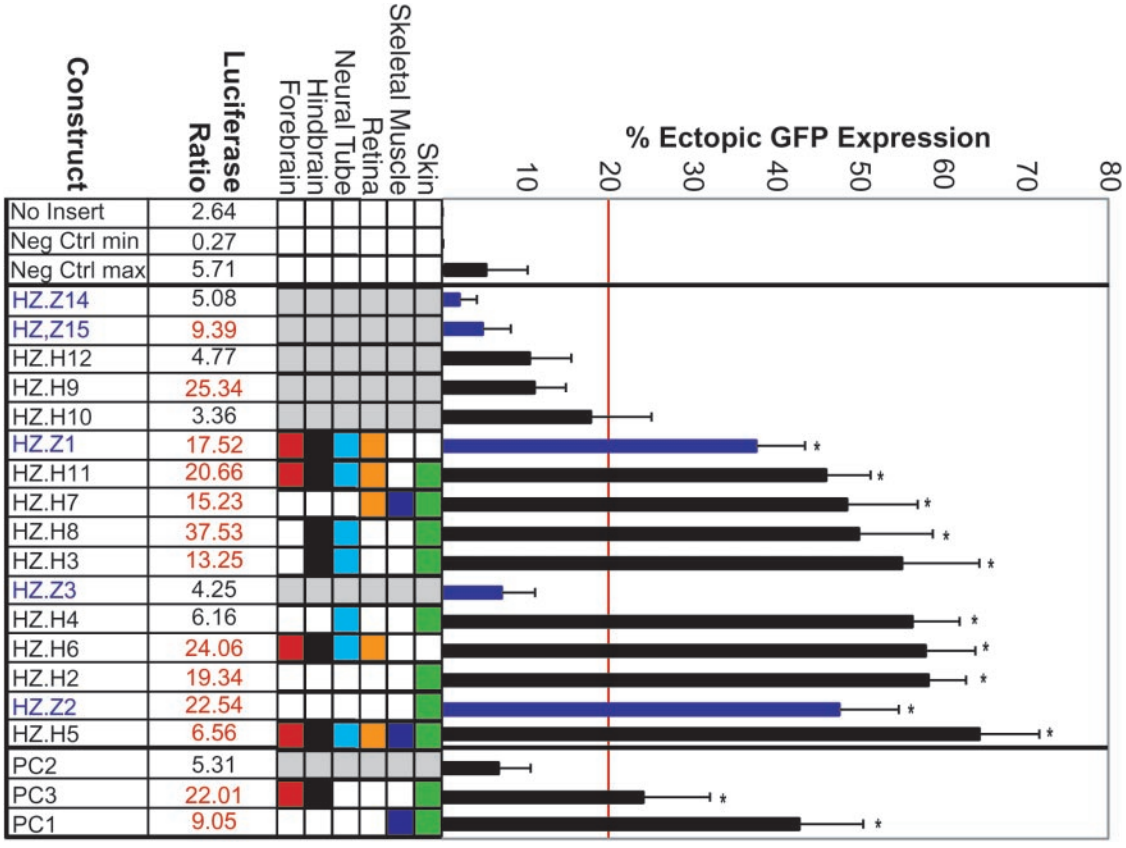

Figure 3. Quantitative analysis of reporter expression demonstrates that HZ NCECR regulate transcription in transgenic zebrafish. Sixteen conserved NC elements (11 from human [black], 1 from zebrafish [blue] and 2 paralogous sequences from both species), as well as negative control (Neg Ctrl, $n = 8$) and positive control (PC, $n = 3$) sequences were cloned for analysis. The percentage of embryos exhibiting extra-cardiac GFP expression are plotted in the horizontal axis (with standard error bars); the red vertical line indicates an approximate threshold delineating positive results which are also marked by an asterisk (*). For GFP-positive sequences, tissues with significant fluorescence in non-cardiac tissues, those tissues/organs which reproducibly demonstrated ectopic fluorescence are denoted with colored boxes. In those embryos where ectopic fluorescence did not exceed the statistical threshold for significance, no tissues are highlighted (gray boxes). Quantiative assessment of luciferase activity is presented in the LR column. LR values shown in this figure are the mean values of at least three samples. Experimental LRs which exceeded the mean of the pooled negative control LR value by 2 SD are highlighted in red.

These sequences were amplified, cloned and analyzed using transient transgenesis in zebrafish. Ectopic expression was observed in multiple tissues including neuronal tissue (forebrain, hindbrain and neural tube), retina, skeletal muscle and skin (e.g. shown in Figure 4). For each NCECR, the tissues containing ectopic expression were highly reproducible (see Figure 3). To examine whether expression modulation by these sequences differed significantly from that seen with negative control sequences, we analyzed the χ^2 metric comparing ectopic fluorescence in the experimental constructs to the pooled set of negative controls with Bonferroni correction for multiple testing. Using these criteria, 10 of 16 injected constructs (63%) demonstrated significant extra-cardiac expression of the GFP reporter (Figure 3). These data suggest that the HZ NCECR dataset is enriched for *cis*-acting transcriptional regulatory sequences.

Quantitation of transcription using dual luciferase analyses in transgenic zebrafish embryos

Transient transgenesis in zebrafish embryos results in mosaic expression and limits the use of single quantitative reporters such as luciferase in the fish. We employed a dual luciferase (firefly and renilla) system to address the issue of heterogeneous expression. Co-injection of two vectors has been shown previously to result in congruent mosaic expression during zebrafish development with this promoter (29). Therefore we reasoned that a two-reporter strategy might control for the heterogeneity of expression associated with zebrafish transient transgenesis. Pooling transgenic embryos were harvested and renilla and firefly luciferase activities were measured independently in each sample. The quotient of the two activities (counts per second) results in a constant luciferase ratio (LR) across samples (data not shown). Thus, the ratio of the luciferase activities serves as a consistent and comparable metric to quantitate expression in zebrafish transient transgenesis.

At 3 dpf, the baseline firefly to renilla LR was 2.75 ± 0.55 following co-injection of two enhancerless vectors. When negative control sequences were included, the average ratio was 2.64 with a SD of 1.88 for the pooled negative control constructs (range 0.27–5.71). We interpreted values which exceeded the mean for these pooled negative control sequences by 2 SD as exerting a significant effect on transcription (threshold ratio = 6.40). Applying this criterion to the positive control enhancers, two elements exceeded the

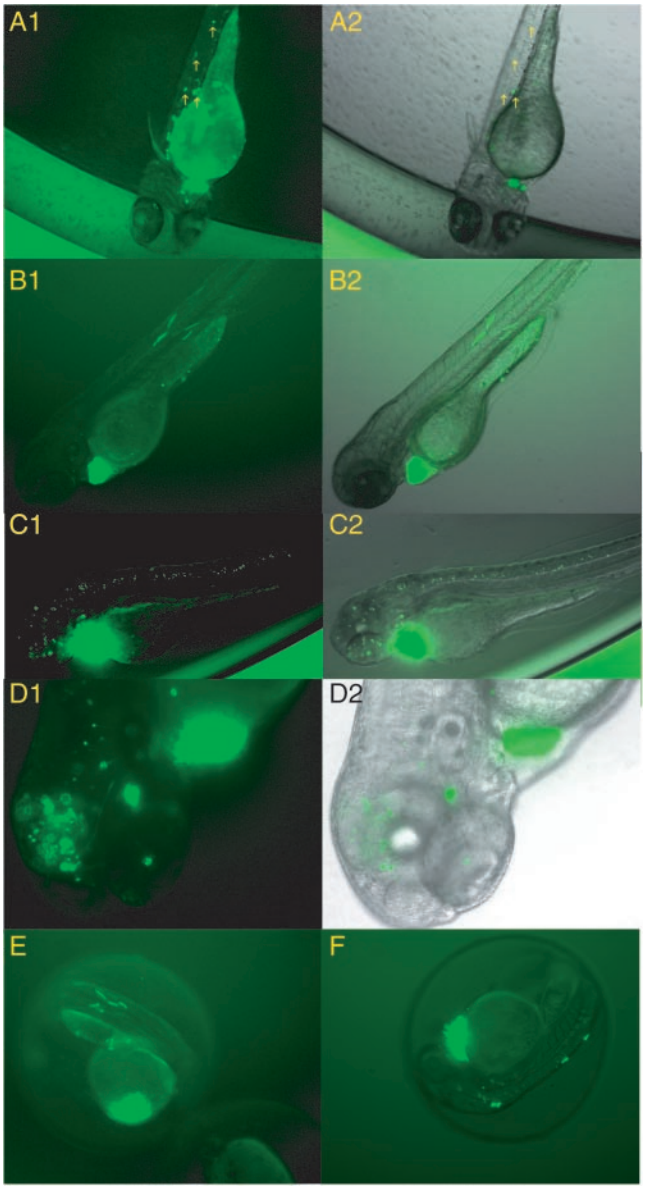

Figure 4. Spatial localization of a GFP expression in transgenic zebrafish. Panels A1–D1 show GFP-images only; panels A2–D2 are merged GFP and bright-field images to assist in the spatial localization of the fluorescent signal. **(A)** GFP expression construct containing conserved sequence HZ.H2 demonstrates ectopic fluorescence in the dermal cells (yellow arrows). **(B)** Skeletal myocytes (elongated cells) and skin fluorescence (round cells) are identified in a 3 dpf transgenic embryo injected with the expression construct containing the HZ.H7 conserved sequence. **(C)** Brain, retina and neural tube fluorescence following injection with HZ.H11:cMLC2:GFP construct. In each case the bright fluorescence between the yolk sac and head represents intrinsic, positive control GFP expression in cardiac myocytes (indicated by the red asterisk in panels A1–C1) directed by the cMLC2 promoter. **(D)** A higher power photomicrograph identifies axonal projections (magnified view shown in the inset, with yellow arrows) and central neuronal expression of GFP in the forebrain and hindbrain of day 3 developing embryos transgenic for the HZ.H6 conserved NC element. Panel **E** and **F** show early (<30 hpf) expression of reporter genes in the skeletal myocytes (panel E) and dermal cells (panel F) prior to hatching.

threshold LR value (PC3 was 22.01 and PC1 was 9.05). PC2 did not have a significant LR (5.31). The positive and negative LR results mirror the spatial enhancement data from the GFP analysis.

Expression constructs were created combining a cloned NCECR, the cMLC2 promoter and firefly luciferase (Figure 1). These were co-injected with an equal amount of the plasmid pGM:RL which contained the Gateway insertion sequence, cMLC2 promoter and RL coding sequence, respectively. Of the 16 anonymous NC sequences tested, 11 (69%) were positive for transcriptional upregulation (LR range: 6.56–37.53). Negative sequences ranged between a minimum value of 3.36 to a maximum of 6.16 (Figure 3).

Spatial regulation and absolute transcription level are concordant but not identical

Comparing the two assays of transcriptional regulation we find that 13 out of 16 NC sequences (81%) have concordant results: 9 demonstrate upregulation by both the spatial and luminometric methodologies and 4 demonstrate no augmentation in both systems. Of the remaining 3 that show discordant results, 2 demonstrate upregulation by luciferase but are unchanged by spatial criteria (HZ.H9, HZ.Z15); the last shows significant ectopic GFP expression but does not achieve the threshold LR value (HZ.H4).

DISCUSSION

A major post-genome challenge remains the detailed functional annotation of NC sequences. The *ab initio* prediction of exons and coding sequences from genomic DNA is aided by the specific selection pressures imposed on such sequences, the presence of ORF and other conserved features which can be used as starting points for the prediction of genes. Identification of regulatory domains in NC sequence has no operational analog to gene prediction algorithms. Predictive algorithms for candidate regulatory NC sequences at present are restricted to direct inter-species comparisons. Ultimately, candidate regions identified through any purely *in silico* strategy will require empiric biological testing.

Comparisons between human and rodent genomes identify numerous conserved NC regions (~60 per predicted gene) (1). However, the large number of genes (~25 000) precludes systematic experimental analysis of potential regulatory elements in a mammalian *in vivo* system. It is possible to restrict the number of candidate sequences by stipulating greater sequence identity (10), but the biologic specificity of such an approach is unknown. Similarly it is possible to use evolutionary distance between species to rationally refine the candidate element prediction. Using a restricted set of candidate NC regulatory sequences identified through a HP comparison, Woolfe *et al.* (13) recently were able to demonstrate *cis*-acting transcriptional regulatory function in transgenic zebrafish. We undertook a comparison between human and draft zebrafish genomic sequences, which elaborates on the role of conserved intergenic sequences.

A comparison of our HZ dataset to the UC and HP conserved elements (10,25), reveals that while there is overlap among the datasets, our comparison also reveals unique elements not found in other comparative datasets. Large clusters of HZ conserved NC sequences overlie the HSA19q and DACH1 gene deserts described previously in human-rodent comparisons. These two genomic regions are known to contain distant gene regulatory elements conserved in the evolution of

vertebrates (12,31). Furthermore, the enrichment of neighboring genes for developmentally conserved functions and clustering of NC ECRs is consistent with analyses of other comparative datasets and supports the concept of a conserved set of NC elements regulating a core vertebrate transcriptional network (10,30,32). The genes which flank these collections of conserved regions are enriched for developmental and transcriptional function by GO analysis. While this may represent the presence of core elements regulating conserved developmental pathways, we cannot exclude that, by using conservation as a filter, we have simply biased our analysis to exclude more recently evolved pathways or genes and thus have detected only those REs associated with more primitive genes and pathways.

A total of 63% ($n = 3010$) of the HZ comparative set contains unique elements identified neither in the UC nor in the HP comparisons. Thus, despite sharing comparable evolutionary divergence from *Homo sapiens* (~450 Mya), significant genomic size differences between *Danio rerio* (1.7×10^9 bp) and *Fugu rubripes* (3.65×10^8 bp), are also reflected in the amount of conserved NC sequences conserved with respect to human genomic sequence. While previous investigations have assigned transcriptional enhancer function as the evolutionary rationale for conservation of these particular sequences, alternative hypotheses regarding the functional raison d'etre of these elements are also tenable. Conserved elements may be involved in transcriptional silencing, as boundary elements or in epigenetic processes such as maintenance of chromatin structure and packing. While empiric testing of the pilot-sized subsets of conserved elements demonstrates transcriptional regulatory function *in-cis*, high throughput methodologies to identify what function, if any, these elements might confer are required.

In order to characterize potential functions of the HZ conserved elements we studied a subset of identified sequences in assays for *cis*-regulatory effects on transcription. We employed an 870 bp segment of the zebrafish cardiac myosin light chain 2 promoter, which is highly organ-specific in both transient and stable transgenesis. The use of an organ-specific promoter in the spatial assessment of transcription offers practical benefits in a transient assay of transcriptional regulation. First, the consistent early cardiac expression driven by the cMLC2 promoter provides a positive control for transgenesis, which may be variable between embryos. Second, the cMLC2 870 bp promoter on its own results in tightly restricted expression providing a null background in all non-cardiac tissues over which to screen for enhancer activity. Importantly, fragments of this promoter have been previously demonstrated to be capable of directing expression in several non-cardiac tissues, suggesting that the basal machinery for such activity is present in our construct.

We demonstrate that 10 of 16 tested HZ conserved sequences confer the ability to modulate tissue-specific transcription from the cMLC2 promoter. This parallels the activity of such sequence elements tested in other vertebrate models, and demonstrates that a HZ comparison identifies elements with similar activities to that seen in HP comparisons, despite the large proportion of unique elements. Our approach differs from that employed by Wolfe *et al*. to investigate the *cis*-regulatory potential of sequences identified in a HP comparison (13). These investigators employed a technique where PCR amplicons harboring potential regulatory sequences are co-injected with linear reporter constructs. We inject circular plasmid DNA, with conserved sequences integrated 5′ to the promoter into fertilized eggs, an established technique for the analysis of regulatory sequences in both transient and stable transgenesis in the fish (33–35).

While analysis of fluorescence *in situ* affords the opportunity to measure reporter activity over time throughout development, the analysis of GFP expression in transgenic zebrafish has limited throughput due to the careful inspection and multiple focal planes necessary to examine all tissues and organs thoroughly. Furthermore, because of the nature of this assay, relative changes in expression restricted to the heart (either augmentation or attenuation) cannot be meaningfully assessed. To address these weaknesses and to provide a scaleable and robust *in vivo* approach for screening candidate regulatory elements, we used an alternative strategy to quantify expression in zebrafish embryos using luciferase-based reporters. This assay is based on the co-injection of (i) a minimal promoter:reporter vector (cMLC2:firefly luciferase) containing a potential enhancer test fragment together with (ii) a second enhancer-less-promoter:reporter vector (cMLC2:renilla luciferase). The co-injection of an experimental construct together with a control construct, whose reporter activity is detected by a distinct chemistry within the same sample, allows for normalization of reporter activity despite the mosaic expression found in zebrafish transient transgenesis. Our findings demonstrate that the luciferase-based assay not only displays consistent background with multiple control elements, but also exhibits considerable dynamic range with the experimental sequences tested. The assay provides complementary data to the spatial information derived from the GFP constructs, in the detection of *cis*-regulation of transcription in transgenic zebrafish embryos. Eleven of 16 experimental sequences showed significant modulation of the firefly:renilla luciferase ratio, including 8 of the 10 unique sequences identified in the GFP spatial assay described above. Finally, the luciferase methodology offers the opportunity to assess reporter activity in an unsupervised fashion in a microtitre plate with an integrated assessment of control activity. This affords the ability to examine the regulatory impact of NCECR with accelerated throughput.

The GFP and luciferase assays are concordant, but not identical, reflecting the differences between the systems despite the fact that the two methodologies share the same promoter. All three positive controls and 87% of the experimental sequences are concordant between the two assays. Of the three discordant cases two had significant LRs but were not significant by the fluorescent, spatial assay (UH9 and UZ15). We interpret this to reflect increased cardiac expression with weak or absent alteration of tissue specificity from the cMCL2 promoter. In the last case where spatial expression was significantly modulated but not significant alteration of the LR was detected, we speculate that the increase in extra-cardiac expression may have been offset by diminished cardiac expression leading to no net increase in overall firefly luciferase activity. While not observed in the set of sequences presented here, measurement of luciferase activity also presents the opportunity to identify negative regulators of transcription (e.g. silencers), which may not be readily evident in any GFP based assay. Thus the quantitation of luciferase

activity provides another 'axis' for the measurement of transcription not possible using fluorescent reporters.

In summary, we have demonstrated that a novel HZ comparison identifies NC elements not present in other distant vertebrate comparisons and confirm that these elements exhibit *cis*-regulatory properties using two different reporter methodologies. These data suggest that annotating functional elements and ultimately mapping gene regulatory networks will require systematic comparisons across multiple species and comprehensive *in vivo* biological validation. The utility of the zebrafish sequence for such large-scale analyses will only be accentuated by the completion of its genome project. In addition, the ability to undertake quantitative expression analyses in high throughput will further enhance the construction of informative models of regulatory networks.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Len Pennacchio, Ed Rubin, Marcelo Nobrega and James Noonan for their helpful discussions and thoughtful commentary. We also thank Kerstin Jekosch Howe at the Sanger Institute's Zebrafish Sequencing Project for her thoughts regarding the manuscript. Research conducted at the Cardiovascular Research Center, MGH was supported in part by grants from the National Heart, Lung and Blood Institute (NHLBI)/National Institutes of Health (NIH) R24HL70580 and 5T32HL007208 and an American College of Cardiology Foundation /Pfizer Postdoctoral Fellowship in Cardiovascular Medicine (JTS). Research conducted at the E.O. Lawrence Berkeley National Laboratory, was partially supported by Grant #HL88728 under the Programs for Genomic Application (NHLBI) and performed under United States Department of Energy Contract DE-AC0378SF00098 to the University of California.

Conflict of interest statement. None declared.

REFERENCES

1. Waterston,R.H., Lindblad-Toh,K., Birney,E., Rogers,J., Abril,J.F., Agarwal,P., Agarwala,R., Ainscough,R., Alexandersson,M., An,P. *et al.* (2002) Initial sequencing and comparative analysis of the mouse genome. *Nature*, **420**, 520–562.
2. Ovcharenko,I., Stubbs,L. and Loots,G.G. (2004) Interpreting mammalian evolution using Fugu genome comparisons. *Genomics*, **84**, 890–895.
3. Barton,L.M., Gottgens,B., Gering,M., Gilbert,J.G., Grafham,D., Rogers,J., Bentley,D., Patient,R. and Green,A.R. (2001) Regulation of the stem cell leukemia (SCL) gene: a tale of two fishes. *Proc. Natl Acad. Sci. USA*, **98**, 6747–6752.
4. Loots,G.G., Locksley,R.M., Blankespoor,C.M., Wang,Z.E., Miller,W., Rubin,E.M. and Frazer,K.A. (2000) Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science*, **288**, 136–140.
5. Pennacchio,L.A., Olivier,M., Hubacek,J.A., Cohen,J.C., Cox,D.R., Fruchart,J.C., Krauss,R.M. and Rubin,E.M. (2001) An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. *Science*, **294**, 169–173.
6. Frazer,K.A., Sheehan,J.B., Stokowski,R.P., Chen,X., Hosseini,R., Cheng,J.F., Fodor,S.P., Cox,D.R. and Patil,N. (2001) Evolutionarily conserved sequences on human chromosome 21. *Genome Res.*, **11**, 1651–1659.
7. DeSilva,U., Elnitski,L., Idol,J.R., Doyle,J.L., Gan,W., Thomas,J.W., Schwartz,S., Dietrich,N.L., Beckstrom-Sternberg,S.M., McDowell,J.C. *et al.* (2002) Generation and comparative analysis of approximately 3.3 Mb of mouse genomic sequence orthologous to the region of human chromosome 7q11.23 implicated in Williams syndrome. *Genome Res.*, **12**, 3–15.
8. Dermitzakis,E.T., Reymond,A., Lyle,R., Scamuffa,N., Ucla,C., Deutsch,S., Stevenson,B.J., Flegel,V., Bucher,P., Jongeneel,C.V. *et al.* (2002) Numerous potentially functional but non-genic conserved sequences on human chromosome 21. *Nature*, **420**, 578–582.
9. Dermitzakis,E.T., Reymond,A., Scamuffa,N., Ucla,C., Kirkness,E., Rossier,C. and Antonarakis,S.E. (2003) Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). *Science*, **302**, 1033–1035.
10. Bejerano,G., Pheasant,M., Makunin,I., Stephen,S., Kent,W.J., Mattick,J.S. and Haussler,D. (2004) Ultraconserved elements in the human genome. *Science*, **304**, 1321–1325.
11. Spitz,F., Gonzalez,F. and Duboule,D. (2003) A global control region defines a chromosomal regulatory landscape containing the HoxD cluster. *Cell*, **113**, 405–417.
12. Nobrega,M.A., Ovcharenko,I., Afzal,V. and Rubin,E.M. (2003) Scanning human gene deserts for long-range enhancers. *Science*, **302**, 413.
13. Woolfe,A., Goodson,M., Goode,D.K., Snell,P., McEwen,G.K., Vavouri,T., Smith,S.F., North,P., Callaway,H., Kelly,K. *et al.* (2005) Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.*, **3**, e7.
14. Ward,A.C., McPhee,D.O., Condron,M.M., Varma,S., Cody,S.H., Onnebo,S.M., Paw,B.H., Zon,L.I. and Lieschke,G.J. (2003) The zebrafish spi1 promoter drives myeloid-specific expression in stable transgenic fish. *Blood*, **102**, 3238–3240.
15. Xu,C., Wu,G., Zohar,Y. and Du,S.J. (2003) Analysis of myostatin gene structure, expression and function in zebrafish. *J. Exp. Biol.*, **206**, 4067–4079.
16. Shentu,H., Wen,H.J., Her,G.M., Huang,C.J., Wu,J.L. and Hwang,S.P. (2003) Proximal upstream region of zebrafish bone morphogenetic protein 4 promoter directs heart expression of green fluorescent protein. *Genesis*, **37**, 103–112.
17. Kimura-Yoshida,C., Kitajima,K., Oda-Ishii,I., Tian,E., Suzuki,M., Yamamoto,M., Suzuki,T., Kobayashi,M., Aizawa,S. and Matsuo,I. (2004) Characterization of the pufferfish Otx2 *cis*-regulators reveals evolutionarily conserved genetic mechanisms for vertebrate head specification. *Development*, **131**, 57–71.
18. Son,O.L., Kim,H.T., Ji,M.H., Yoo,K.W., Rhee,M. and Kim,C.H. (2003) Cloning and expression analysis of a Parkinson's disease gene, uch-L1, and its promoter in zebrafish. *Biochem. Biophys. Res. Commun.*, **312**, 601–607.
19. Luo,W., Williams,J., Smallwood,P.M., Touchman,J.W., Roman,L.M. and Nathans,J. (2004) Proximal and distal sequences control UV cone pigment gene expression in transgenic zebrafish. *J. Biol. Chem.*, **279**, 19286–19293.
20. Her,G.M., Yeh,Y.H. and Wu,J.L. (2004) Functional conserved elements mediate intestinal-type fatty acid binding protein (I-FABP) expression in the gut epithelia of zebrafish larvae. *Dev. Dyn.*, **230**, 734–742.
21. Lin,C.Y., Chen,Y.H., Lee,H.C. and Tsai,H.J. (2004) Novel *cis*-element in intron 1 represses somite expression of zebrafish myf-5. *Gene*, **334**, 63–72.
22. Wang,T.M., Chen,Y.H., Liu,C.F. and Tsai,H.J. (2002) Functional analysis of the proximal promoter regions of fish rhodopsin and myf-5 genes using transgenesis. *Mar. Biotechnol. (NY)*, **4**, 247–255.
23. Tidyman,W.E., Sehnert,A.J., Huq,A., Agard,J., Deegan,F., Stainier,D.Y. and Ordahl,C.P. (2003) *In vivo* regulation of the chicken cardiac troponin T gene promoter in zebrafish embryos. *Dev. Dyn.*, **227**, 484–496.
24. Huang,C.J., Tu,C.T., Hsiao,C.D., Hsieh,F.J. and Tsai,H.J. (2003) Germ-line transmission of a myocardium-specific GFP transgene reveals critical regulatory elements in the cardiac myosin light chain 2 promoter of zebrafish. *Dev. Dyn.*, **228**, 30–40.
25. Ovcharenko,I., Nobrega,M.A., Loots,G.G. and Stubbs,L. (2004) ECR Browser: a tool for visualizing and accessing data from comparisons of multiple vertebrate genomes. *Nucleic Acids Res.*, **32**, W280–W286.
26. Karolchik,D., Baertsch,R., Diekhans,M., Furey,T.S., Hinrichs,A., Lu,Y.T., Roskin,K.M., Schwartz,M., Sugnet,C.W., Thomas,D.J. *et al.*

(2003) The UCSC Genome Browser Database. *Nucleic Acids Res.*, **31**, 51–54.
27. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al.* (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–261.
28. Rozen,S. and Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.*, **132**, 365–386.
29. Rottbauer,W., Saurin,A.J., Lickert,H., Shen,X., Burns,C.G., Wo,Z.G., Kemler,R., Kingston,R., Wu,C. and Fishman,M. (2002) Reptin and pontin antagonistically regulate heart growth in zebrafish embryos. *Cell*, **111**, 661–672.
30. Boffelli,D., Nobrega,M.A. and Rubin,E.M. (2004) Comparative genomics at the vertebrate extremes. *Nat. Rev. Genet*, **5**, 456–465.
31. Kim,J., Bergmann,A., Lucas,S., Stone,R. and Stubbs,L. (2004) Lineage-specific imprinting and evolution of the zinc-finger gene ZIM2. *Genomics*, **84**, 47–58.
32. Ovcharenko,I., Loots,G.G., Nobrega,M.A., Hardison,R.C., Miller,W. and Stubbs,L. (2005) Evolution and functional classification of vertebrate gene deserts. *Genome Res.*, **15**, 137–145.
33. Stuart,G.W., McMurray,J.V. and Westerfield,M. (1988) Replication, integration and stable germ-line transmission of foreign sequences injected into early zebrafish embryos. *Development*, **103**, 403–412.
34. Stuart,G.W., Vielkind,J.R., McMurray,J.V. and Westerfield,M. (1990) Stable lines of transgenic zebrafish exhibit reproducible patterns of transgene expression. *Development*, **109**, 577–584.
35. Gilmour,D.T., Jessen,J.R. and Lin,S. (2002) In Nusslein-Volhard,C. and Dahm,R. (eds), Manipulating gene expression in the zebrafish. *Zebrafish: A Practical Approach*. Oxford University Press, Oxford, Vol. 261, pp. 121–143.