Cracking the genome's second code: Enhancer detection by combined phylogenetic footprinting and transgenic fish and frog embryos

Miguel L. Allende [a,*], Miguel Manzanares [b], Juan J. Tena [c], Carmen G. Feijóo [d], José Luis Gómez-Skarmeta [c]

[a] *Millennium Nucleus in Developmental Biology, Facultad de Ciencias, Universidad de Chile, Casilla 653, Santiago, Chile*
[b] *Instituto de Investigaciones Biomédicas and Centro de Biología Molecular Severo Ochoa, CSIC-UAM, Madrid, Spain*
[c] *Centro Andaluz de Biología del Desarrollo, Consejo Superior de Investigaciones Científicas and Universidad Pablo de Olavide, Sevilla, Spain*
[d] *Facultad de Ciencias de la Salud, Universidad Nacional Andrés Bello, Santiago, Chile*

Abstract

Genes involved in vertebrate development are unusually enriched for highly conserved non-coding sequence elements. These regions are readily detected *in silico*, by genome-wide sequence comparisons between different vertebrates, from mammals to fish (phylogenetic footprinting). It follows that sequence conservation must be the result of positive selection for an essential physiological role. An obvious possibility is that these conserved sequences possess regulatory or structural functions important for gene expression and, thus, an *in vivo* assay becomes necessary. We have developed a rapid testing system using zebrafish and *Xenopus laevis* embryos that allows us to assign transcriptional regulatory functions to conserved non-coding sequence elements. The sequences are cloned into a vector containing a minimal promoter and the GFP reporter, and are assayed for their putative *cis*-regulatory activity in zebrafish or *Xenopus* transgenic experiments. Vectors used include plasmid DNA and the Tol2 transposon system in fish and *X. laevis*. We have followed this logic to detect and analyze conserved elements in an intergenic region present in the *Iroquois* (*Irx*) gene clusters of zebrafish, *Xenopus tropicalis*, *Fugu rubripes* and mouse. We have assayed ~50 of these conserved elements and shown that the majority behave as modular positive regulatory elements (enhancers) that contribute to specific temporal and spatial domains that are part of the endogenous gene expression pattern. Moreover, comparison of the activity of cognate *Irx* enhancers from different organisms demonstrates that conservation of sequence is accompanied by *in vivo* functional conservation across species. Finally, for some of the most conserved elements, we have been able to identify a critical core sequence, essential for correct enhancer function.

Keywords: Enhancer; Transgenic; Regulatory sequence; Iroquois; Tol2

1. Introduction

Analysis of the control of gene expression involves the identification of *cis*-regulatory sequences, both positive (enhancer) and negative (silencer), elements that are bound by *trans*-activating or repressive factors that modulate transcription. Traditionally, these elements are identified by laborious deletion analysis of genomic sequences that surround the gene of interest (promoter bashing) followed by testing different fragments in transfected cells using reporter genes. Often, it is possible to detect binding sites for regulatory proteins by sequence analysis of the regions near the transcriptional start site of the gene, as the binding sites have consensus motifs. However, *cis*-elements may be hundreds of kilobases away from the gene, making it difficult to define the appropriate region in which to perform a search. Moreover, as transcription factors usually recognize sequences of only a few base pairs, it is likely that many spurious elements will be incorrectly identified [27]. Thus, to reveal authentic *cis*-acting elements, it is essential to

[*] Corresponding author. Fax: +562 276 3802.
E-mail address: allende@uchile.cl (M.L. Allende).

confirm their activity in a system that recapitulates as closely as possible the endogenous situation. In this sense, *in vivo* systems are much preferred over cell culture systems where all temporal and spatial information on transcriptional regulation is lost.

The technique termed phylogenetic footprinting takes advantage of the availability of genomic sequences of different species to search for homologous regions that may harbor regulatory elements [16,17]. The phylogenetic distance between the compared species is proportional to the degree of conservation of regulatory elements, as is the case for coding sequences. However, since non-coding regulatory sequences have fewer constraints in terms of position and can be dispersed over long genomic distances, they are more challenging to discover. Specialized software has been developed to help identify these sequences.

Surprisingly, examination of the sequenced genomes of vertebrates has revealed numerous highly conserved non-coding regions (HCNRs; [21,22,18]). This is true even among distantly related species, such as fish and mammals, whose last common ancestor existed over 3–400 Myr[1] ago. The existence of HCNRs suggests functional importance, as they are often more conserved than the coding sequences of protein-coding genes. Our ability to detect HCNRs with computational methods is due to the fact that they are of considerable length, usually 100–500 bps. Therefore, they could either have a structural role, regulating for example, chromatin accessibility or nuclear matrix attachments, or be *cis*-regulatory regions that concentrate binding sites for multiple factors. Alternatively, both functions could also co-exist in the same HCNR. Genomic analysis of HCNRs has revealed, unexpectedly, that they are prevalent in the vicinity of developmentally important genes [18,6]. These genes are conserved throughout the animal kingdom in terms of sequence and function and it is likely that the regulatory networks that govern their expression are conserved as well.

The *Iroquois* (*Irx*) genes are a group of homeodomain-containing transcription factors that participate in neural patterning during embryonic development [8]. Originally identified in *Drosophila* [10], they are present in all vertebrates in differing numbers depending on the species. Mammals have two clusters (*IrxA* and *IrxB*) of three genes each [9,13]. Fish, on the other hand, have more genes due to an extra whole genome duplication at the base of the teleost lineage [14]. There are 10 *Irx* genes in *Fugu rubripes* and 11 in *Danio rerio* [11,12]. Interestingly though, in all species of vertebrates, the clustered structure of the *Irx* genes is generally maintained suggesting a functional requirement for keeping the chromosomal disposition of the genes intact. This led us recently to carry out an analysis of a large intergenic region spanning the distance between the clustered *IrxB* genes in a search for functional regulatory elements [1]. To achieve this goal, we carried out comparisons between the *Irx3–Irx5* and *Irx5–Irx6* intergenic regions of several vertebrate species and we found numerous conserved sequences. We isolated ~50 of these from zebrafish and *Xenopus tropicalis* and cloned them in a plasmid vector that contains a basal zebrafish or *X. tropicalis Irx3* promoter and the gene encoding the Enhanced Green Fluorescent Protein (EGFP). Injection of these constructs into zebrafish or *Xenopus* embryos allowed us to assay whether the isolated sequences behaved as enhancers of transcription. Our results demonstrate that this is a simple and quick method for validating the functionality of potential enhancer elements *in vivo*. In addition, we have more recently introduced the use of the Tol2 transposon system for transient transgenesis in fish, a method that reduces the mosaicism of EGFP expression inherent to other methods [15]. We envision that comparative genomics coupled with rapid functional testing in fish or frog embryos will be a convenient tool for annotating sequence elements involved in transcriptional regulation in vertebrate genomes.

2. Methods and results

2.1. *Care and raising of animals, transgenesis protocols*

Zebrafish of the AB wild type strain were kept at our own facility on a 14–10 h light–dark cycle under standard conditions [7]. Embryos were obtained by natural spawning of adult fish and were incubated in petri dishes at 28 °C in E3 medium (5 mM NaCl, 0.17 mM KCl, 0.33 mM $CaCl_2$, 0.33 mM $MgSO_4$, and 0.1% methylene blue). Embryonic stages [19] are expressed as hours post-fertilization (hpf). For microinjection, embryos were collected immediately after fertilization, mounted in injection chambers without removal of the chorion and microinjected with pulled glass capillaries held by a micromanipulator and attached to a microinjector (MPPI-2 Pressure Injector System, World Precision Instruments). Injection chambers are of two types. In the first case, we pour molten agarose (1% in water) into a 5-cm plastic petri dish. Prior to solidification, a glass slide is inserted into the agarose at an angle such that it generates a gentle slope ending in a vertical edge. The dish is filled with E3 medium and embryos are placed along the edge of the injection ramp; this allows them to be punctured by the injection needle through the chorion and injected. In the second method, a glass microscope slide is set into a 10 cm plastic petri dish and embryos (with chorions) are deposited with a transfer pipette along the edge of the slide. Excess water is removed once enough embryos are in the dish and they are injected as above. The embryos are held against the edge of the slide by surface tension and exposure to air is not detrimental if injection time is less than 30 min per dish. Once all embryos are injected, E3 medium is added and the entire dish is placed in an incubator at 28 °C.

[1] *Abbreviations used:* Myr, millions of years; kb, kilobases; EGFP, enhanced green fluorescent protein.

Delivery of DNA and RNA can be done in embryos up to the four-cell stage without significant hindrance to distribution of the nucleic acids. We normally inject 3–5 nl of solution into the yolk cell as molecules can freely diffuse to the animal pole blastomeres until the 32-cell stage [20]. We inject linearized plasmid DNA at a concentration of 50–100 ng/μl, which is near the toxicity tolerance of the embryos (300–500 pg/embryo). With plasmid DNA, it is important to use the highest possible concentration to reduce the degree of mosaicism; this is particularly important when the expression pattern of the transgene is restricted to a small region of the embryo. In our hands, about 50% of the plasmid-injected embryos survive to 24 hpf without major developmental defects; embryos showing malformations are excluded from subsequent analysis.

For Tol2 transposon [15] injections, we inject a mix containing circular DNA at a concentration of 5 ng/μl and transposase mRNA at 25 ng/μl. Since we inject between 5 and 10 nl in the yolk, we are injecting each embryo with 25–50 pg of DNA and 125–250 pg of mRNA. Survival rates were higher (75%) than those obtained with plasmid DNA injections, and embryos were healthier, as less DNA was injected.

All microscopic observations of embryos older than 20 hpf were done after anaesthetizing in MS222 (3-aminobenzoic acid ethyl ester, methanesulfonate salt, Sigma). EGFP-expressing embryos were photographed with a digital CCD camera (Magnafire, Optronix) mounted on an MZ-12 dissecting microscope (Leica). Images were handled with Photoshop 7.0 for Macintosh.

Xenopus laevis were purchased from Xenopus Express, France. The animals were kept at our own facility at 19 °C on a 12–12 h light–dark cycle under standard conditions. Females were induced to ovulate by injecting them with Human Chorionic Gonadotropin hormone into the dorsal lymph sac (500–800 units). Females begin to lay eggs about 9–10 h after induction of ovulation. To minimize the risk of diseases during this period, females are kept isolated in clean water with 20 mN NaCl and 5 μg/ml of gentamycin from the moment of treatment until 24 h later. Eggs are collected manually with a smooth abdominal massage. To generate transgenic embryos, we followed the protocol of Amaya and Kroll [28] with the recent modification described in Sparrow et al. [29]. This protocol relies in the integration of foreign DNA into sperm nuclei that is subsequently used for egg fertilization. The modification described in Sparrow et al. [29], avoids the requirement of sperm nuclei chromatin decondensation mediated by egg extracts and the use of restriction enzymes to facilitate DNA integration. In addition, sperm nuclei are prepared as described in Amaya and Kroll [28] except that permeabilization is done with 100 μg/ml digitonin. In brief, in this simplified method, 250,000 sperm nuclei in 2.5 μl are incubated at room temperature for 15 min with linear DNA (100–150 ng) in 2.5 μl of water. A reaction aliquot (2.1 μl) is then diluted into 200 μl of sperm dilution buffer (250 mM sucrose; 75 mM KCl; 0.5 mM spermidine trihydrochloride; 0.2 mM spermine tetrahydrochloride, pH 7.3–7.5; [28]) and mixed several times by pipetting up and down with a clipped tip to generate a homogeneous suspension. This suspension is then loaded into injecting needles that are prepared by pulling 30 μl Drummond micropipettes and clipping the end at a 30–40 μm diameter. Collected eggs are dejellied with 2.2% cysteine hydrochloride in 1× MMR (adjusted to pH 7.9 with NaOH) and washed several times in 1× MMR (10× Marc's Modified Ringer's (MMR): 1 M NaCl; 20 mM KCl; 10 mM $MgCl_2$; 20 mM $CaCl_2$; 50 mM HEPES, pH 7.5). Eggs are then transferred into 1% agarose-coated dishes containing 0.4 × MMR with 6% Ficoll and 50 μg/ml gentamycin. Nuclear transplantation is performed, with an infusion syringe pump (Harvard Apparatus model 22) and a 2.5 ml Hamilton glass tight syringe. This allow a continuous flow of 10 nl/s. When the embryos have reached the 4-cell stage, they are separated from uncleaved eggs and transferred into a separate agarose-coated dish containing 0.1 × MMR + 6% Ficoll + 50 μg/ml gentamycin. In our hands, this transgenic method allowed us to obtain an efficiency of about 20% of the embryos that survive to gastrulation. Transgenic *X. laevis* embryos were examined for EGFP fluorescence with a fluorescence-dissecting microscope from gastrula to tadpole stages. After this stage, the embryos were fixed for *in situ* hybridization analyses with an EGFP riboprobe. An enhancer was considered active when it promotes the same pattern in at least five different embryos. It should be noted that each embryo corresponds to a different integration event.

Protocols involving animals have been reviewed by the Animal Welfare and Ethics Committees of our institutions, the University of Chile and Universidad Pablo de Olavide.

2.2. Comparative genomic analysis and isolation of selected genomic regions

Genomic sequences spanning the entire *IrxB* cluster where downloaded from the Ensembl genome server (www.ensembl.org), using the latest releases. The comparative analysis was re-checked each time new releases of available whole genome sequences or new organisms were incorporated into the database. In this way, we have compared and used (at some point in our analysis) sequences from eight different vertebrates species, including three mammals (human, *Homo sapiens*; mouse, *Mus musculus*; and rat, *Rattus norvegicus*), one avian (chicken, *Gallus gallus*), one amphibian (*X. tropicalis*) and three teleost fishes (zebrafish, *Danio rerio*, and two pufferfishes, *Takifugu rubripes* and *Tetraodon nigroviridis*). Occasionally, the presence or absence or certain sequences was double-checked in databases for un-assembled genomic sequences from the species under study. The genomic sequences of the single *Iroquois* cluster from *Drosophila melanogaster* and the two pairs of *Irx* genes from the urochordate *Ciona intestinalis* ([24] and unpublished results) were also compared with selected vertebrate species, showing no obvious sequence conservation apart from the protein-coding regions of the *Irx* genes. A preliminary search for the presence of vertebrate conserved sequences in draft genomic fragments of the cephalochordate amphioxus *Branchiostoma floridae* revealed no similarities, although we did identify at least two different *Irx* genes in this organism (unpublished results).

Sequence comparisons of the *Irx* clusters between different species were performed using both local and global alignment algorithms, using the servers and tools freely available at the Pipmaker ([25]; http://bio.cse.psu.edu/pipmaker) and Vista ([26]; http://genome.lbl.gov/vista) sites, respectively. In the case of local alignments, all positions in one sequence are compared with all in the second sequence in both orientations, while in global alignments the order and orientation of similar sequence regions are maintained. The powers and limitation of both approaches have been discussed elsewhere [17,27], and when starting a comparative genomic project it is convenient to use both tools. Results obtained for multi-species comparison of the vertebrate *Irx* clusters showed the same results using both tools, and for simplicity further analysis was performed with the Vista set of tools using default settings, because of its more user-friendly web interface.

Once conserved regions were identified and selected, primers were designed to flank the conserved sequences; the primers were placed at a distance of approximately 100–200 nucleotides on each side of the region identified by Vista in the displayed alignments. PCRs were carried out using 100 pg of genomic DNA obtained from our stocks of wild type zebrafish, *X. tropicalis*, *Fugu rubripes* (MRC Geneservice) or mouse. Before transferring the PCR-amplified fragments to the expression vector (see below), they were inserted by TA cloning into the pGEM-Easy plasmid (Promega), which contains two *Eco*RI sites flanking the PCR insert.

2.3. Vector construction and delivery

As the majority of selected conserved fragments lie far from the presumed genes they regulate, we decided to test their activity when coupled singly with a basal promoter from one of those genes. This approach attempts to recapitulate with a single enhancer the endogenous situation in which enhancer-bound proteins interact with basal transcription factors and RNA polymerase and recruit them to promoters. By definition, an enhancer should be able to function independent of distance, orientation and promoter. Nonetheless, we are aware that the strategy we have employed will preclude us from identifying positive regulatory elements that are strictly dependant on their structural organization with respect to the basal promoter, that show promoter specificity, or that have repressor function.

To select an adequate zebrafish promoter, we examined the sequence of the immediate upstream region of the *Irx3a* gene and we arbitrarily designed primers spanning 0.6 Kb of DNA 5′ of the ATG codon; we assumed that any important promoter elements would be contained in such a fragment. It was important for the ensuing experiments, to ascertain that this fragment of DNA lacked any transcriptional activity on its own. We therefore cloned the *Irx3a* promoter in front of the EGFP gene (isolated from the pGreenLantern vector; Gibco) in a pBluescript backbone generating plasmid *pzIrx3a-GFP*. Injection of this construct in linearized form into zebrafish embryos did not produce any GFP expression, confirming that the *Irx3a* promoter is silent in the absence of enhancer elements. To test the potential enhancer activity of individual elements isolated from zebrafish, we cloned each isolated fragment into *pzIrx3a-EGFP* using an *Eco*RI site upstream of the *Irx3a* promoter. If the genomic sequence contained internal *Eco*RI sites, we used the *Sac*II and *Pst*I sites flanking the insert in the pGEMT-Easy vector for transferring the conserved region into the *pzIrx3a-EGFP* plasmid. Each construct was then linearized with *Xho*I and tested *in vivo* by microinjection and analysis of expression of the EGFP reporter during embryogenesis.

For testing putative enhancer elements in transgenic *X. laevis* embryos, we carried out a parallel strategy. In this case, we generated the *pXIrx3-EGFP* plasmid containing 0.6 Kb of the *Xenopus Irx3* basal promoter combined in turn with each genomic fragment. The *Xenopus*-conserved sequences were inserted 5′ to the promoter in the *pXIrx3-EGFP* plasmid using the *Eco*RI sites from the PGEMT vector. If the genomic sequence contained internal *Eco*RI sites, we did not use the *pXIrx3-EGFP* plasmid. Instead, we used the *pzIrx3a-EGFP* expression vector and the *Sac*II or *Pst*I sites. These enhancer constructs were also linearized with *Xho*I. To determine whether promoter specificity was an issue with any of the previously identified enhancers, we also used alternative basal promoters in experiments using transgenic *Xenopus* embryos. We cloned the enhancers as before using, in addition to zebrafish and *Xenopus Irx3* promoters, the opsin promoter from *Xenopus*. In the case of the enhancer analyzed in this fashion, all three promoters gave similar results in *X. laevis* transgenics [1]. However, it should be mentioned that not all enhancers behave similarly with different promoters. In some cases, we detected transcriptional activity only when the enhancer was combined with the species-specific *Irx3* promoter.

Subsequent to our initial work with the *Iroquois* enhancers [1], we have continued with this approach but have modified the transgenesis technique for zebrafish. We are now using the Tol2 transposable element system [15] for generating transient transgenic fish. This has involved cloning the DNA of interest (in our case, the genomic fragment to be tested, the basal promoter and the EGFP gene) into the Tol2 vector (pT2KXIG), which contains two terminal inverted repeats flanking the cloning site. The inverted repeats are recognized by the Tol2 transposase, which must be supplied separately. Thus, we co-inject transposase-encoding mRNA together with the circular DNA vector into one-cell stage zebrafish embryos see [15]. The Tol2 transposase is synthesized and the enzyme catalyzes the recombination event that will integrate the DNA into the host genome. Importantly, integration is very efficient, which results in embryos that are far less mosaic than when plasmids are used (Compare Fig. 1A, B with C). A second advantage of transposons is that each integration event involves a single copy of the exogenous DNA. In addition, the high efficiency of the transposon system increases the chances of germline integration of the foreign DNA, therefore simplifying the generation of stable transgenic lines. Contrariwise, linear plasmids present the inconvenience of concatemerization after injection and typically integrate as multiple copies [2], generating heterogeneity in the expression levels of the transgenes. Moreover, plasmids may cause genomic rearrangements at the integration site, with the risk of introducing mutations.

2.4. Analysis of transient transgenic animals

Zebrafish embryos injected with expression plasmids containing the different conserved genomic fragments were raised to the appropriate developmental stage and scored for expression of EGFP using a dissecting microscope equipped with epi-fluorescence. Each fragment was injected into 300–400 embryos of which about 50% survived the first day of development. Of these, usually about 10% expressed EGFP though the expression level and the location varied with each enhancer. Within a batch of embryos injected with the same enhancer, expression was consistent in terms of tissue specificity (Fig. 1A). We normally observed a few cells expressing EGFP in random locations throughout the embryo, but these never represented more than 5% of the total number of expressing cells. By cataloguing the distribution of EGFP-positive cells from 10 to 15 embryos, it was possible to establish the expression domain conferred by each enhancer. This type of analysis for transient transgenic fish expressing a reporter in mosaic fashion has been used by other authors [3–5].

When the Tol2 system is used in zebrafish, the degree of mosaicism drops considerably. While plasmid injections result in EGFP expression in about 10% of cells within a tissue or expression domain, the same experiment done with a Tol2-based vector consistently results in 80% of cells expressing the reporter (Fig. 1B). This degree of penetrance approaches that obtained in stable transgenic zebrafish embryos (Fig. 1C) or that obtained by using transgenic *Xenopus* embryos (Fig. 1D), where mosaicism is non-existent.

In the *X. laevis* transgenic protocol, the exogenous DNA is integrated into the male genome prior to fertilization. As there is no mosaicism in these animals, only a few surviving embryos are sufficient to obtain enough data to establish the expression domain of the enhancer. However, in some cases expression of EGFP could not be easily visualized under fluorescent illumination due to the opacity of the embryonic tissues. Therefore, we resorted to performing *in situ* hybridization using an EGFP-specific probe to detect expressing tissues (Fig. 1D).

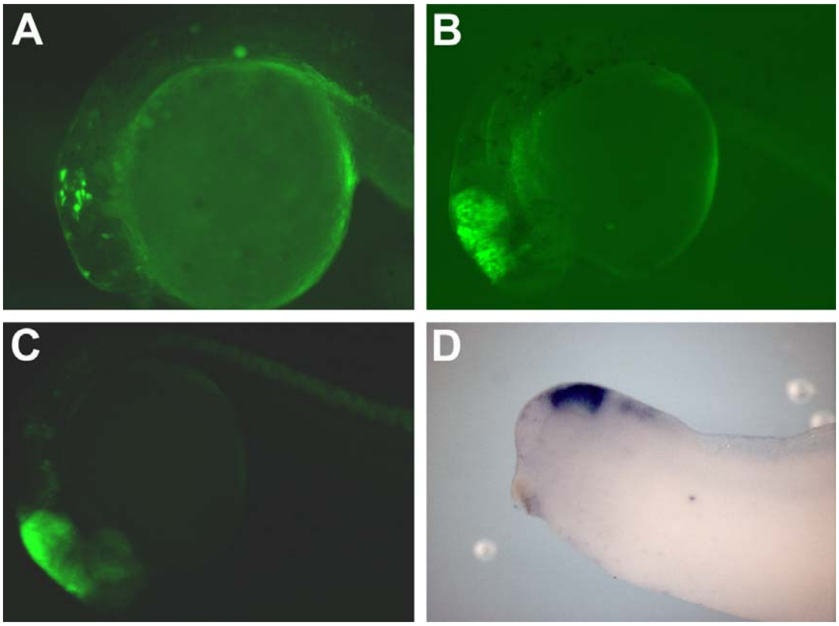

Fig. 1. Testing enhancer elements in transgenic fish and frog embryos. (A) Zebrafish embryos were injected at the one-cell stage with plasmid *pz54390Irx3aGFP* and raised to 24 hpf. Of an injected batch of embryos, only about 20% present GFP expression, and these typically show mosaic expression of the reporter. Nonetheless, it is possible to observe that the *z54390* enhancer element drives expression of GFP restricted to the midbrain. (B) Injection of *z54390Irx3aGFPTol2* transposable element shows expression in most cells of the dorsal midbrain. In a typical experiment, more than 50% of the surviving embryos show this type of strong expression. Moreover, weakly expressing transposon-injected embryos were always less mosaic than even the strongest expressing embryos injected with plasmid DNA. (C) This panel shows a stable transgenic embryo (F1) resulting from integration of the *z54390Irx3aGFPTol2* vector. Note the uniform (non-mosaic) expression of GFP in the midbrain and compare to (A and B). (D) A *Xenopus laevis* transgenic embryo injected with linearized *pz54390Irx3aGFP* plasmid DNA, fixed and treated by *in situ* hybridization to detect GFP mRNA expression. Note that the embryo is non-mosaic revealing the entire expression domain of the enhancer element.

In our study, we tested 23 and 26 conserved regions isolated from the zebrafish and *Xenopus* genomes, respectively. From these, 16 from zebrafish and 10 from *Xenopus* showed enhancer activity. We also tested eight elements corresponding to homologous sequences shared between zebrafish and *Xenopus*. However, while all eight zebrafish sequences were active, only four of the *Xenopus* ones promoted EGFP expression. In the cases where both zebrafish and *Xenopus* homologous sequences showed enhancer activity, we did not observe the exact same expression pattern in both species. These differences may be due to small variations in the DNA sequences of orthologous HCNRs. It is also possible that some of the HCNRs have evolved in specific vertebrate lineages to acquire distinct functional properties. A third alternative is that the dissimilar activities are a consequence of the different techniques used for the generation of transgenic zebrafish and *Xenopus*. Finally, the transparency of zebrafish embryos may help to detect expression in more domains in zebrafish than is possible in *Xenopus*. Accordingly, we found many more active elements in zebrafish than in *Xenopus*.

3. Discussion and conclusions

Several studies have made clear the importance of comparative genomics for the identification of transcriptional regulatory elements in the genomes of vertebrates. Powerful new bioinformatic tools have been key in the detection of these sequences and the next step involves the analysis of their function. As many of the genes harboring HCNRs in their vicinity are involved in development, it is important to test these in a developmental setting, using organisms in which embryos are plentiful, easily accessible, and amenable to gene transfer technology. Fish and frog embryos satisfy these requirements and they have already been used in several studies involving functional enhancer profiling.

The main difficulty in this endeavor is to correctly identify the putative enhancers from genomic sequences. While the prediction of the protein-coding portions of genes is now straightforward due to specific differences in the base pair composition of these regions and other sequence hallmarks such as intron–exon boundaries, this is not so for *cis*-regulatory elements. Consensus sequences for transcription factor binding sites have been used for prediction of regulatory elements, but their short length and often loose stringency can lead to an over-estimation of putative versus real sites of 1–1000 [27]. More recently, other approaches such as co-occurrence of various different transcription factor binding sites, or motif search in genomic regions of co-expressed genes have been developed that more accurately detect functional regulatory elements [27].

The advent of whole genome sequences for a variety of vertebrate species now allows narrowing down the region in which to look for putative regulatory elements. The extensive sequence comparisons between species has been termed phylogenetic footprinting. The logic behind this approach is simple. Sequences that are conserved during evolution above a certain threshold must exist due to positive selective pressure for their retention. This is the case for protein-coding regions of genes, and therefore some functional role must exist for other conserved sequences that do not code for proteins or that are not part of transcribed regions of the genome. These functions are not restricted to *cis*-acting transcriptional regulation, but can also include other functions such as chromatin structure and assembly.

The choice of species to use in a comparative genomic approach will depend greatly on the region and genes under study. In some cases, comparison of human and mouse sequences will be enough to detect specific conserved regions among large areas of non-conservation. However in other cases, such as the *Irx* clusters, the degree of conservation between mammals is so high that it renders these comparisons ineffective. The inclusion of more distantly related species allows the filtering out of sequences until a discrete number is retained that can then be assayed for a functional role.

In the case of the *Irx* clusters, mammalian versus fish sequence comparisons have identified an adequate number of conserved non-coding sequence elements that could then be tested functionally. The high degree of evolutionary conservation of the function of this gene family during vertebrate development supports the hypothesis that putative regulatory elements identified this way will represent fundamental components of the regulatory networks controlling the precise spatial and temporal regulation of *Irx* genes in the developing embryo.

We have developed a quick assay system to test predicted enhancer elements for functional activity *in vivo*. This approach is particularly useful for developmentally important genes, as observation of the expression patterns is straightforward in fish and frog early embryos. Injection of constructs containing the putative enhancers coupled to a minimal promoter and GFP produces reproducible expression of the reporter in both systems. In the case of the enhancers detected from within the *Irx* intergenic regions, we were able to partially recapitulate the endogenous expression domains with the isolated elements suggesting that we are able to dissect the transcriptional regulation of this complex developmental regulator [1]. The data generated in this way will eventually be fed into the information database that will construct gene regulatory networks in the different vertebrates.

The use of linearized plasmid-based DNA constructs for the generation of transient or stable transgenic animals has been predominant over the years, due to the minimal manipulation needed after a simple cloning step. This approach was originally devised for the production of mouse transgenics where it is still widely used. In *X. laevis*, plasmid DNA integrates into the sperm nuclear genome at reasonably high frequency. The “modified” sperm can then be used for fertilization of wild type eggs. As DNA can integrate prior to fertilization, some of the resulting embryos are transgenic, and since they are not chimaeric, they show proper spatial and temporal regulation of integrated promoter constructs. This strongly facilitates the analysis of the precise expression pattern promoted by the integrated DNA since, in contrast to transgenesis in mice or zebrafish, breeding of the animals is not required. In fish embryos, this technique has not been accessible and linearized injection of plasmids in embryos presents several drawbacks. The injected DNA rapidly forms concatemers and becomes integrated relatively late in the genome of only a few cells. Thus, the degree of mosaicism is high, expression is variable and many embryos must be injected to obtain a convincing picture of the localization of reporter gene expression. One alternative method to naked plasmid injections is to use the *Sce*I meganuclease protocol [23]. This technique involves injecting a plasmid that contains two *Sce*I restriction sites flanking the transgene of interest. The enzyme is co-injected with the plasmid, which produces a linear transgene and facilitates nuclear translocation. Transgenic frequency is high, mosaicism low and integration appears to be mostly as single copies.

Another alternative to plasmid injections in zebrafish has been the introduction of an efficient transposable element system, the Tol2 system [15]. We show here that our constructs inserted in the Tol2 vector produce more consistent expression as the foreign DNA integrates very efficiently early in development, likely as single copies. This advance brings the fish up to par with respect to transient transgenic analysis in *Xenopus*, with the added benefit of embryo transparency, indispensable for precisely mapping tissue-specific expression domains. The choice of EGFP as a reporter is critical for facilitating the identification of expression profiles when little or nothing is known about the enhancer specificity. Live embryos can be observed at all timepoints, and early onset of expression is easily detected. In addition, and in contrast to mouse, in zebrafish and *Xenopus* it is possible to examine the enhancer activity of a particular region in the same embryo throughout development. One possible drawback is that EGFP protein accumulation is slow and may not precisely reflect the timing of transcriptional activity of the enhancer. Therefore, it is good practice to confirm fluorescent protein detection with immunostaining or *in situ* hybridization against the product of the reporter gene.

Since adequate and cheap technology is available for enhancer profiling in both fish and frog embryos, which system should we choose? If possible, it is best to use both, as long as the features under analysis are conserved between them. If a putative enhancer behaves in a similar fashion and directs expression to equivalent domains in both species, we can make a much more reliable prediction of its function. It must be remembered that fish and frogs are just as evolutionarily distant as fish and mammals are. Using both models reduces the possibility of misjudgment of the expression profile of an enhancer. Furthermore, when comparing sequences across-species, it may be relevant to pinpoint not only the similarities but also the differences. These may be defining features that generate the variations in expression domains or timings that occur in different organisms, and in such cases the availability of both models can help to pinpoint the regulatory origin of these differences.

The method we have described here promises to be of use in assigning regulatory function to HCNRs. However, we are aware of the limitations of such an approach and can clearly identify two aspects of transcriptional regulation that will require certain modifications to this approach. First, our method is oriented towards the identification of enhancer elements but is ineffective when examining silencer or repressor elements. It will be necessary to modify the vectors we have used to allow for detection of repressive activity, ideally by using a weak promoter that is expressed ubiquitously. In second term, the enhancer elements we have tested have been completely removed from their endogenous genomic context. It is well established that chromatin structure is critical for the correct regulation of gene activity and that rearrangements can be spread over hundreds of kilobases. It is very likely that some of the conserved elements that turned up negative in our assay require a chromatin context that was not reproduced when placed in isolation. These elements will be much more difficult to study and it may even be necessary to take a loss-of-function approach to determine their role.

Acknowledgments

We acknowledge the technical help of Catalina Lafourcade and Florencio Espinoza. We thank Koichi Kawakami for providing the pT2KXIG vector. M.A. and C.G.F. were supported by Fondecyt (1031003), ICM (P02-050), ICGEB (CRP/CHI03-03c) and UNAB (10-03); M.M. by the Spanish Ministry of Education and Science and the EMBO YIP Programme. J.L. from Spanish Ministry of Education and Science (BFU2004-00310) and Junta de Andalucía and M.A. and J.L.G-S by a Universidad de Chile-CSIC collaboration grant (2004CL0018).

References

[1] E. de la Calle-Mustienes, C.G. Feijóo, M. Manzanares, J.J. Tena, E. Rodríguez-Seguel, A. Leticia, M.L. Allende, J.L. Gómez-Skarmeta, Genome Res. 15 (2005) 1061–1072.

[2] G.W. Stuart, J.V. McMurray, M. Westerfield, Development 103 (1988) 403–412.

[3] F. Müller, B. Chang, S. Albert, N. Fischer, L. Tora, U. Strahle, Development 126 (1999) 2103–2116.

[4] P. Blader, C. Plessy, U. Strahle, Mech. Dev. 120 (2003) 211–218.

[5] T. Dickmeis, C. Plessy, S. Rastegar, P. Aanstad, R. Herwig, F. Chalmel, N. Fischer, U. Strahle, Genome Res. 14 (2004) 228–238.

[6] C. Plessy, T. Dickmeis, F. Chalmel, U. Strahle, Trends Genet. 21 (2005) 207–210.

[7] M. Westerfield, The Zebrafish Book: A Guide for the Laboratory Use of the Zebrafish, *Danio rerio*, University of Oregon Press, Eugene, OR, 2000.

[8] J.L. Gómez-Skarmeta, J. Modolell, Curr. Opin. Genet. Dev. 12 (2002) 403–408.

[9] A. Bosse, A. Stoykova, K. Nieselt-Struwe, K. Chowdhury, N. Copeland, N.A. Jenkins, P. Gruss, Dev. Dyn. 218 (2000) 160–174.

[10] C. Dambly-Chaudière, L. Leyns, Int. J. Dev. Biol. 36 (1992) 85–91.

[11] R. Dildrop, U. Ruther, Dev. Genes Evol. 214 (2004) 267–276.

[12] C.G. Feijóo, M. Manzanares, E. de la Calle-Mustienes, J.L. Gómez-Skarmeta, M.L. Allende, Dev. Genes Evol. 214 (2004) 277–284.

[13] T. Peters, R. Dildrop, K. Ausmeier, U. Ruther, Genome Res. 10 (2000) 1453–1462.

[14] J.S. Taylor, Y. Van de Peer, I. Braasch, A. Meyer, Philos. Trans. R. Soc. Lond. B Biol. Sci. 356 (2001) 1661–1679.

[15] K. Kawakami, H. Takeda, N. Kawakami, M. Kobayashi, N. Matsuda, M. Mishina, Dev. Cell 7 (2003) 133–144.

[16] F. Müller, P. Blader, U. Strahle, BioEssays 24 (2002) 564–572.

[17] K.A. Frazer, L. Elnitski, D.M. Church, I. Dubchak, R.C. Hardison, Genome Res. 13 (2003) 1–12.

[18] A. Woolfe, M. Goodson, D.K. Goode, P. Snell, G.K. McEwen, T. Vavouri, S.F. Smith, P. North, H. Callaway, K. Kelly, et al., PLoS Biol. 3 (2005) e7.

[19] C.B. Kimmel, W.W. Ballard, S.R. Kimmel, B. Ullmann, T.F. Schilling, Dev. Dyn. 203 (1995) 253–310.

[20] C.B. Kimmel, R.D. Law, Dev. Biol. 108 (1985) 78–85.

[21] G. Bejerano, M. Pheasant, I. Makunin, S. Stephen, W.J. Kent, J.S. Mattick, D. Haussler, Science 304 (2004) 1321–1325.

[22] A. Sandelin, P. Bailey, S. Bruce, P.G. Engstrom, J.M. Klos, W.W. Wasserman, J. Ericson, B. Lenhard, BMC Genomics 5 (2004) 9.

[23] V. Thermes, C. Grabher, F. Ristoratore, F. Bourrat, A. Choulika, J. Wittbrodt, J.S. Joly, Mech. Dev. 118 (2002) 91–98.

[24] S. Wada, M. Tokuoka, E. Shoguchi, K. Kobayashi, A. Di Gregorio, A. Spagnuolo, M. Branno, Y. Kohara, D. Rokhsar, M. Levine, H. Saiga, N. Satoh, Y. Satou, Dev. Genes Evol. 213 (2003) 222–234.

[25] S. Schwartz, Z. Zhang, K.A. Frazer, A. Smit, C. Riemer, J. Bouck, R. Gibbs, R. Hardison, W. Miller, Genome Res. 10 (2000) 577–586.

[26] K.A. Frazer, L. Pachter, A. Poliakov, E.M. Rubin, I. Dubchak, Nucleic Acids Res. 32 (2004) W273–W279, Web Server issue.

[27] W.W. Wasserman, A. Sandelin, Nat. Rev. Genet. 5 (2004) 276–287.

[28] E. Amaya, K.L. Kroll, Methods Mol. Biol. 97 (1999) 393–414.

[29] D.B. Sparrow, B. Latinik, T.J. Mohun, Nucleic Acids Res. 28 (2000) E12.