Understanding the regulatory genome

M. EVA ALONSO1,2,3, BÁRBARA PERNAUTE1,2, MIGUEL CRESPO1,2, JOSÉ LUIS GÓMEZ-SKARMETAP3, and MIGUEL MANZANARES*,1,2

1Instituto de Investigaciones Biomédicas CSIC-UAM, Madrid, 
2Department of Cardiovascular Developmental Biology, Centro Nacional de Investigaciones Cardiovasculares (CNIC), Madrid and 
3Centro Andaluz de Biología del Desarrollo CSIC-UPO, Sevilla, Spain

ABSTRACT The sequencing of the whole genome of multiple species provides us with the instruction book of how to build an organism and make it work, plus a detailed history of how diversity was generated during evolution. Unfortunately, we still understand only a small fraction, which is locating where genes are and deciphering the proteins they code for. The next step is to understand how the correct amount of gene products are produced in space and time to obtain a fully functioning organism, from the egg to the adult. This is what is known as the regulatory genome, a term coined by Eric H. Davidson. In this review, we examine what we know about gene regulation from a genomic point of view, revise the current in silico, in vitro and in vivo methodological approaches to study transcriptional regulation, and point to the power of phylogenetic footprinting as a guide to regulatory element discovery. The advantages and limitations of each approach are considered, with the emerging view that only large-scale studies and data-crunching will give us insight into the language of genomic regulatory systems, and allow the discovery of regulatory codes in the genome.

KEY WORDS: transcriptional regulation, comparative genomics, transgenesis

Introduction

Any cell is constantly receiving multiple different inputs from its external environment, by means of different signalling mechanisms, and subsequently has to respond according to its lineage and location. One of the primary ways to do so is by controlling the level of expression of different genes. Such changes in gene expression will result in changes in the proteins and RNAs present in the cell and therefore in a phenotypical response to the stimulus.

Regulation of gene function occurs at many different levels. DNA is transcribed within the nucleus of the cell into an RNA molecule, which in turn can then be subjected to splicing. Following, protein-coding mRNAs are translated into peptide sequences in the cytoplasm. The protein is then folded into a 3-dimensional protein structure that will determine its function. Regulation of gene function happens at every single one of these steps. Chromatin can be epigenetically marked to allow or not a certain DNA segment to be available for transcription. The process of epigenetic modification of DNA is strictly controlled and a crucial point in regulation of cellular phenotypes (Kouzarides, 2007).

The next point of regulation would be transcriptional regulation per se: the control of when, where and how much transcript is produced from a given gene. This includes the assembly of the general transcription machinery onto the basal or proximal promoter, and is influenced by other sequence elements located elsewhere (see below). Furthermore, the existence of alternative promoters and their differential use brings in an added level of complexity. Once the mature RNA is exported to the cytoplasm, further points of regulation are in place, both in the control of splicing and in the translation of mRNAs into proteins. And finally, proteins themselves can be subject to modifications that will determine their function and stability.

Significant progress has been made in all of these aspects in recent years, but by and large it has become clear that the regulation of the transcriptional activity of a gene is the most studied of these control points. This is due, in part, to the availability of robust experimental approaches to study transcriptional regulation, and also to the hope that we are near to deciphering the code and language used in the process. Understanding the paradigm of differential gene expression will be fundamental in order to know how organisms develop, differentiate, and tackle a constantly changing environment. Thanks to the availability of full genome sequences from multiple metazoan
species (Fig. 1), we are closer to unravel this paradigm.

The Book of Life, Chapter 2

The metaphor of the genome as the book of life has been extensively used. In fact, at a click of the mouse, we have access to the full and complete set of instructions for building an organism. This is at the same time fascinating and deeply frustrating, as we can read the instructions but have not yet been able to fully understand them, and hence are not able yet to reconstruct from scratch even the simplest creature (Lartigue et al., 2007). Nevertheless, a huge effort has been made to at least understand part of these instructions, and at present the fraction of the genome that is made of peptide-coding exons has been characterized in detail and is practically completed. So, in the first chapter of the book, we have learned which the parts of the machine are. The next task is to unravel how to combine them into a functioning organism.

The coding part of genes if immersed in random DNA will be basically inactive. The protein they encode will be invisible to selection if other sequence motifs and proteins capable of regulating its function did not exist. Genes will then necessarily be flanked by regulatory sequences that, together with the expression and activity of proteins encoded elsewhere, regulate their function under determined environmental conditions in specific cells or tissues. Regulatory sequences therefore are as important for gene function as the coding sequences that determine the amino acids composition of a protein (Wray et al., 2003).

The next goal in genome biology will be to identify all regulatory sequences that control the function of each gene and to understand how the input of these elements is coordinated to execute complex cellular processes. We are beginning to see the results of such efforts, where all functional elements from selected regions of the human genome are being characterized and described (ENCODE Project Consortium, 2007). Such global and systematic approaches will surely serve as the base for a full understanding of how is life’s complexity encoded by genomes.

The regulation of gene transcription

By regulating the transcription of a gene, it is possible to control the amount of functional product available, be it an RNA molecule
or a protein. Traditionally, it has been considered that transcription is regulated through the interplay between proximal and distal sequence elements located in a cis configuration (that is, on the same molecule, or in this case, chromosome) that will determine if the gene must be transcribed or maintained silent. The basal promoter, where the general transcriptional machinery interacts, is the main proximal element. Although multiple efforts have been devoted to define what constitutes a promoter, and genome-wide location of promoters in mammals have been performed (Carninci et al., 2006), it is still too often to find loose descriptions of these in many studies. By taking the proximal one or two kilobases (kb) upstream from the transcriptional start site, many authors assume that they are dealing with the promoter for that gene. If this may surely be true, what is not taken into consideration is that multiple other regulatory elements can be taken on board with such an approach. Therefore, and from here on, we will rather use a functional definition of a promoter as a directional sequence element that is necessary but not sufficient for the transcription of a gene, located in close proximity to its 5’ end, and that will necessarily need other regulatory elements to effectively direct transcription.

Distal elements are usually classified into enhancers, which will activate transcription from the promoter, and silencers, which will negatively regulate the activity of the promoter. Due to the technical approaches used for their study, it has been much easier to identify and characterize positive (enhancer) elements. The study of negative (silencer) elements requires a more complex and sophisticated experimental design and is difficult to separate from other aspects such as gene silencing by chromatin modifications (Sengupta et al., 2004).

The classical definition states that an enhancer is a regulatory element located in cis that activates transcription from a promoter independently of its orientation and location. Enhancers are supposedly composed by binding sites for transcription factors that will be recruited to the DNA and then interact with the basal transcriptional machinery located at the promoter. However, we must not forget that these are strictly functional definitions, defined by what we know up to now. The evidence nowadays points to a much more complex situation, where elements can have opposing roles in different situations (positive or negative), or that different levels of regulation (epigenetic, transcriptional, structural) are combined at the same regions of DNA. Therefore, we believe that a more encompassing definition must be made, and hereafter we will refer to all sequence elements that take part in the control of the transcription of a gene as cis-regulatory elements. This broader designation is necessarily less informative, but at the same time will help to have an unbiased look on how gene transcription is regulated.

Where do genes start and end?

Genes are typically identified as the genomic regions that are transcribed into a functional RNA molecule, be it an mRNA that will be translated into a protein or any of the multiple non-protein-coding RNA molecules (miRNAs, piwi-RNAs, snoRNAs, antisense mRNAs) that are being described daily. In this view, genes begin at the transcriptional start site or at the proximal promoter (if it has been characterized in some detail) and end at the polyadenylation site. However, if we define a gene as all the DNA sequences necessary for the production of a functional molecule, the limits start to blur.

Approximately 25% of the human or mouse genome consists of gene-poor regions greater than 500 kb, termed gene deserts (Nobrega et al., 2003). These segments have been minimally explored, and their functional significance remains elusive. One category of functional sequences postulated to lie in gene deserts are regulatory elements. Recent studies have shown that regulatory elements have the ability to modulate gene expression over very long distances (Lettice et al., 2002), what is consistent with gene deserts containing such elements even if their targets genes are located hundreds of kilobases away. In transgenic assays in mice, frogs and zebrafish, several of these regions have been shown to act as cis-regulatory elements (Muller et al., 2002).

The observation that distal genomic region as far as 1 Mb appear to affect the expression of specific genes, such as Sox9, a temporal tissue-specific transcription factor involved in male sexual development and bone formation (Bien-Wilner et al., 2007), points to the possibility of elements lying at great distances from the coding region of genes being necessary to achieve proper expression. According to different models of gene regulation, distal cis-acting regulatory regions may be brought in proximity to the genes of interest by the interaction between transcription factors and other proteins sitting at these sites with protein complexes present at the promoter. In this way, by looping out the intervening DNA, an active chromatin hub or similar structure will be created (Tolhius et al., 2002; Dillon, 2006).

Such views underscore the possible existence of higher-order chromosomal domains containing multiple genes subject to similar and shared regulatory mechanisms. The first evidence for such domains came from the description of the structure and expression of the β-globin and Hoxclusters, where multiple genes that originated by local tandem duplications are orderly expressed in time and space during development as result of their chromosomal organization. In fact, regulatory elements that globally control the expression of these and other gene clusters have been identified in vertebrates (Montoliu et al., 1996, Spitz et al., 2003). Even if these examples refer to genes belonging to the same family, where a scenario for the evolution of common regulatory control is not too difficult to imagine, recent evidence shows that this could be a more general phenomenon.

The existence of co-expression domains in the genome, where unrelated genes are expressed in a similar fashion (Hurst et al., 2004), points to global regulatory mechanisms acting over broad chromosomal regions enclosed by insulator elements. These act as barriers to protect genes from both positive and negative influences of their genomic or chromatin environment and thus maintain the accurate temporal and spatial transcriptional patterns critical to normal development (Yoon et al., 2007). Two types of insulators have been defined (Kuhn and Geyer, 2003; Capelson and Corces, 2004; Gaszner and Felsenfeld, 2005; Yoon et al., 2007). One of them acts as barrier to protect genes from chromosomal position effects by preventing the spread of heterochromatin-mediated silencing. The other type, enhancer-blocking insulators, protects promoters from activation by distal enhancers. The majority of them contain a highly conserved consensus motif for the CTCF (CCTC-binding factor) factor (Wallace and Felsenfeld, 2007). Genome-wide mapping of such sites appear to define these global domains (Kim et al., 2007). On
the other hand, the analysis of synteny blocks in vertebrate genomes, that are regions where gene order has been conserved through evolution, also can be used as a guide for the detection of global regulatory domains in the genome (Kikuta et al., 2007).

Finally, there is longstanding evidence for regulation not in cis but in trans, such as the transvection phenomena in Drosophila (Duncan, 2002), where sequence elements located on other chromosomes can affect gene expression (Chen et al., 2002). In yeast, tRNA genes, although dispersed along the genome, are brought together in the nucleus to be transcribed co-ordinately, revealing the necessity of a complex three-dimensional organization of the genome for proper regulation of gene expression (Thompson et al., 2003). Such cases show that the complexity of the problem we are aiming to study is daunting, and that the comprehensive understanding of how the expression of any given gene is properly regulated may be an impossible task with our current tools. However, keeping this in mind, at least we are able to start understanding the basics of transcriptional regulation.

**How can we study cis-regulatory elements? The language of transcriptional regulation**

The activity of cis-regulatory elements is largely mediated by the sequence-specific binding of transcription factors to the DNA molecule. If we know the sequence of the binding site for a specific transcription factor, the identification of regulatory elements under its control would be straightforward. Unfortunately, transcription factor binding sites cannot reliably be identified from sequences comparison alone (Carey and Smale, 2000). Although sequence scans can identify candidate binding sites, confirmation that a particular sequence motif actually functions in regulation of gene expression (Table 1).

![Table 1](https://example.com/table1.png)

<table>
<thead>
<tr>
<th>Method</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>in silico</strong></td>
<td>TF binding site prediction</td>
<td>very low predictive value</td>
</tr>
<tr>
<td></td>
<td>quick approximation</td>
<td>extremely high false positive rate</td>
</tr>
<tr>
<td>EMSA</td>
<td>rapid and inexpensive</td>
<td>no cellular context</td>
</tr>
<tr>
<td></td>
<td>biochemical evidence for binding of TF to DNA</td>
<td>possible artificial results</td>
</tr>
<tr>
<td></td>
<td>reagents (oligos, proteins) readily available</td>
<td></td>
</tr>
<tr>
<td><strong>in vitro</strong></td>
<td>Tissue culture</td>
<td>no spatio-temporal resolution</td>
</tr>
<tr>
<td></td>
<td>rapid and inexpensive</td>
<td>forced ectopic expression can lead to false-positive results</td>
</tr>
<tr>
<td></td>
<td>some approximation to tissue specificity</td>
<td>no discrimination on reg. element type (e., + or -)</td>
</tr>
<tr>
<td>ChIP-on-chip</td>
<td>cellular chromatin context intact</td>
<td>only indicates DNA occupancy, not reg. element activity</td>
</tr>
<tr>
<td></td>
<td>whole-genome scale studies</td>
<td>no spatio-temporal resolution</td>
</tr>
<tr>
<td>ChIP-seq</td>
<td>Zebrfish</td>
<td>no deletion by homologous recombination</td>
</tr>
<tr>
<td></td>
<td>external development</td>
<td>genetic redundancy, due to telost genome duplication</td>
</tr>
<tr>
<td></td>
<td>easy rearing through complete life cycle</td>
<td>little use for the study of mammalian specific characters</td>
</tr>
<tr>
<td></td>
<td>high transgenic efficiency</td>
<td></td>
</tr>
<tr>
<td></td>
<td>reporter gene imaging in live embryos and adults</td>
<td></td>
</tr>
<tr>
<td></td>
<td>possible to combine with experimental embryology</td>
<td></td>
</tr>
<tr>
<td></td>
<td>availability of diverse genetic reagents</td>
<td></td>
</tr>
<tr>
<td><strong>in vivo</strong></td>
<td>Xenopus</td>
<td>costly rearing through complete life cycle</td>
</tr>
<tr>
<td></td>
<td>external development</td>
<td>little transparency of embryo</td>
</tr>
<tr>
<td></td>
<td>high transgenic efficiency</td>
<td></td>
</tr>
<tr>
<td></td>
<td>possible to combine with experimental embryology</td>
<td></td>
</tr>
<tr>
<td>Chick</td>
<td>amniote system</td>
<td>only transient studies feasible</td>
</tr>
<tr>
<td></td>
<td>relative ease in transgenics by electroperoration</td>
<td>high mosaicism of transgene</td>
</tr>
<tr>
<td></td>
<td>possible to combine with experimental embryology</td>
<td></td>
</tr>
<tr>
<td></td>
<td>mammalian system</td>
<td>costly rearing</td>
</tr>
<tr>
<td></td>
<td>possible to delete reg. elements by homologous recombination</td>
<td>low transgenic efficiency</td>
</tr>
<tr>
<td></td>
<td>availability of diverse genetic reagents</td>
<td>internal development</td>
</tr>
</tbody>
</table>

A significant improvement to in vitro binding assays has been the ChIP-on-chip (chromatin immunoprecipitated samples hybridized to high-density microarrays chips) technique (Ren et al., 2000) and the recent Chip-Seq improvement (Barski et al., 2007). This is a genome-wide location analysis for the isolation and identification of the DNA sequences occupied by specific DNA binding proteins in cells. These binding sites may indicate func-
tions of various transcriptional regulators and help identify their target genes. The identified binding sites may also be used as a basis for annotating functional elements in genomes. All types of \textit{cis}-regulatory elements can be identified using ChIP-on-chip (promoters, enhancers, repressors and silencing elements, insulators, boundary elements, sequences that control DNA replication) but this technique will not allow us to differentiate one from the other. The value of ChIP-on-chip as a global approach to study protein-DNA interactions is illustrated by its use as a technological platform by the ENCODE consortium to map all functional sequence elements in the human genome (ENCODE Project Consortium, 2007). Multiple variations on this approach have been developed (Loh et al., 2006), but once more the information gathered this way will not give us insight on what these putative regulatory elements are actually doing, but only that they are occupied by a DNA-binding factor present in a particular cellular context. Nevertheless, this approach will sure provide extremely valuable data, as it is scaled up to include binding data for every single transcription factor in the genome.

The first glimpse into the regulatory activity of \textit{cis}-regulatory elements can be obtained by studies in tissue culture using reporter assays. In these, the DNA fragment to be tested, linked to a reporter gene, is transfected into a cell line and its activity measured in response to different stimuli or the co-transfection of an expression construct for a candidate gene to regulate such elements. These assays are quick and inexpensive, but in many cases provide only partial information about the physiological role of \textit{cis} elements and their regulation by different transcription factors. One of the main caveats is that the forced expression of a transcription factor from a transfected construct will usually give extremely high levels of expression, outside of the range normally found in vivo. Under such conditions, it might not be surprising that such protein will bind a putative consensus and regulate the expression of the reporter gene. Furthermore, cell lines used in tissue culture can harbour multiple chromosomal abnormalities, making them a poor proxy to what is actually occurring in the organism, and results obtained using different lines are therefore not readily comparable.

Although much has been learned about enhancers and other \textit{cis}-regulatory elements using the approaches outlined above, testing the regulatory activity in the context of the whole organism will ultimately be the only way to resolve tissue specific spatiotemporal expression (Gomez-Skarmeta et al., 2006). Therefore, the generation and analysis of transgenic organisms will be needed if we want to fully comprehend how regulation of gene expression is achieved. In standard practice, each fragment is individually cloned together with a minimal heterologous promoter or the gene’s own promoter and coupled to a reporter gene, such as lacZ or a fluorescent protein. These constructs are then delivered, most usually into early embryos, and the activity of the reporter is observed as readout of the regulatory potential of the fragment. Transgenic animals can be made using many different species, but those most used (at least in vertebrate developmental biology studies) are mouse, \textit{Xenopus}, zebrafish and chick.

Being a mammal, and therefore more closely related and theoretically a better model for human disease, the mouse has most often been used in these assays. DNA is directly microinjected into the one-cell embryo which is allowed to develop to the desired stage (transient transgenics) or bred to obtain a stable transgenic line. The disadvantages of this model are that the manipulation of the mouse is labour-intensive and costly, transgenic efficiency is low, and it is not possible to continuously observe development as embryos need to be dissected for observation. Among the advantages of the mouse is the fact that many different mutant and over-expressing lines are available, which allows performing genetic analyses of regulatory interactions. Despite the limitations mentioned above, large-scale studies of regulatory elements are starting to be performed in mice (Pennacchio et al., 2006). Finally, homologous recombination in ES cells provides the tool for the definitive test of the function of a putative regulatory element: its deletion from the genome and the analysis of the effect of the mutation on gene expression \textit{in vivo}.

The analysis of \textit{cis}-regulatory elements in chick embryos have recently been made possible by in ovo electroporation of DNA constructs (Itasaki et al., 1999). In this way, it is possible to analyze different genomic fragments for their capacities (Uchikawa et al., 2003). The main limitation of the technique is that the electroporated DNA is not stably integrated in the genome, but functions as an episome. This results in a highly mosaic expression of the reporter, and its dilution as the time from electroporation to observation increases. The promise of the use of Tol2 transposon-mediated gene transfer, in a similar way to zebrafish (see below), could help to overcome these problems (Sato et al., 2007).

Unlike mouse and chick, \textit{Xenopus} and zebrafish have the advantage that their development can be continuously observed due to transparency of the egg. If the reporter gene used to follow \textit{cis}-regulatory element activity can be visualized \textit{in vivo} (such as fluorescent proteins), transgenic embryos can be followed and reared to adulthood to obtain a stable line. The main disadvantage of transgenic \textit{Xenopus} is that obtaining stable lines requires prolonged rearing of colonies and the space required for such breeding programmes. These problems are not present in zebrafish, which at the same time is a vertebrate system very useful for genetic analysis. This organism is particularly attractive for this purpose, because females can produce a large number of eggs, embryonic development is rapid, each generation can be produced within 60-70 days, embryos are transparent throughout the early developmental processes, and large numbers of fishes can be raised in a relatively small space. Advances, such as the use of the Tol2 system or the clever design of promoter combinations to test repressor activity (Amsterdam and Becker, 2005), make zebrafish an extremely convenient and efficient system for the large scale analysis of vertebrate \textit{cis}-regulatory elements, situating it as a serious choice when compared to mouse (Allende et al., 2006).

\textbf{Narrowing down: the use of evolution as a tool}

As we have seen, such complex transcriptional regulation is mediated by the coordinated binding of transcription factors to discrete, typically noncoding DNA sequences, allowing the integration of multiple signals to regulate the expression of specific genes. These sequences can be up to several hundred bases in length, although not necessarily composed exclusively by individual transcription factor binding sites, and may be located at distances of several hundred kilobases to over a megabase in
Fig. 2. Phylogenetic footprinting at the Irx8 cluster. The Irx8 cluster is located at the distal portion of mouse chr. 8 and spans more than 1 Mb of an extremely gene-poor region (A). The comparison of a 130 kb region adjacent to mouse Irx3 to different vertebrates (human, chick, Xenopus and pufferfish) shows how only some fragments have been evolutionarily conserved (B). Those highly conserved non-coding elements can then be tested for cis-regulatory activity by transgenesis in different species. In this example (C), the selected fragment (arrow) drives expression of the reporter gene (GFP in zebrafish and Xenopus, and lacZ in mouse) to the pronephros in all three vertebrates. Browser images were downloaded from Ensembl (A, www.ensembl.org) and the ECR Browser (B) from dcode.org (ecrbrowser.dcode.org).
either direction from the genes on which they act (Bishop et al., 2000). Moreover, these fragments may not act on the closest gene but can act across intervening genes (Spitz et al., 2003) and can also be located within neighbouring genes (Lettice et al., 2003). This leaves us with the conundrum of where to look for regulatory elements when studying the transcriptional regulation of a gene. Being aware that we will be missing relevant information, and taking into account the limitations of current assays in hand to analyze regulatory elements, we must limit in some way the available search space (that, theoretically, would encompass the whole genome).

The first step in this direction is to limit the genomic region surrounding the gene to study. A useful rule of thumb is to assume that the majority of regulatory elements will be located in the vicinity of the gene, taking as limits the neighbouring 5′ and 3′ genes as long as their expression does not show significant overlap with the gene under study. If such overlap exists, we must assume the possibility of common regulatory mechanisms acting on more than one gene (see above), and consequently extend the region to search. This approach will leave us with anything from some kilobases (that can be easily scanned) to megabases (Nobrega et al., 2003), that will still need further narrowing down.

A powerful approach for the identification of putative regulatory elements has been the comparison of genomic sequences between species, what has been termed phylogenetic footprinting (Muller et al., 2002; Frazer et al., 2003). The utility of comparative sequence analysis is based on the hypothesis that important biological sequences are evolutionarily conserved between species due to functional constrains. An obvious case is that of peptide-coding exons, but others are not as easily explained. Therefore, evolution is helping us out by highlighting those regions in the genome that for some reason, have been conserved.

Examination of the sequenced genomes of vertebrates (Fig. 1) has revealed numerous highly conserved non-coding regions, even between distantly related species that diverged more than 350 Mya, such as fish and mammals, (Bejarano et al., 2004; Sandelin et al., 2004; Woolfe et al., 2005). The detection of highly conserved sequence elements by computational methods is feasible because of their considerable length (100-500pbs) and over 70% similarity (Boffelli et al., 2004). Nevertheless, we should keep in mind that these are arbitrary values, tuned to discriminate between those regions that are clearly conserved and other that are not and with little biological significance on their own (see below). Such elements could have a structural role, for example controlling chromatin accessibility or nuclear matrix attachments, or be cis-regulatory regions that concentrate binding sites for multiple factors. It has been shown that these highly conserved elements are preferentially located in the vicinity of genes coding for transcription factors involved in early development, the trans-dev set (Plessy et al., 2005, Woolfe et al., 2005). 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 (Davidson, 2006).

The power of phylogenetic footprinting to dissect the transcriptional regulation over complex genomic regions is clearly illustrated by the study of the vertebrate Iroquois/Irx complexes (Fig. 2). The Irx genes code for homeodomain transcription factors that participate in multiple steps of pattern formation during embryonic development (Cavodeassi et al., 2001). Originally identified in Drosophila they are conserved throughout the animal kingdom and organized in clusters both in Drosophila and in vertebrates. In mammals, the two Irx clusters (IrxA and IrxB) span over 1 Mb of DNA, with no other genes in between. Therefore, they are located in what is a prime example of a gene desert and a clear case where conventional enhancer-bashing by deletion analysis is not viable. The presence of shared and global cis-regulatory element is surely one of the reasons for the evolutionary conservation of such peculiar organization (Duboule, 1998).

The Irx clusters are among the genomic regions with highest content in evolutionarily conserved non-coding sequences (Sandelin et al., 2004, Woolfe et al., 2005), and it is only when mammals versus fish genomic comparison are used that a number of elements that is feasible to test in vivo are identified (de la Calle-Mustienes et al., 2005). A systematic survey of these regions has been carried out in Xenopus and zebrafish transgenics for the IrxB cluster, demonstrating the presence of multiple cis-regulatory elements that drive reporter expression in overlapping domains (de la Calle-Mustienes et al., 2005). Furthermore, by comparing the ability of a conserved element to drive expression in the same domain in different species (Fig. 2), the correlation between conservation of sequence and conservation of function can be tested.

Powers and pitfalls of phylogenetic footprinting as a guide for cis-element identification

As we have seen above, a significant number of evolutionarily conserved non-coding regions are functional but, in contrast, some other apparently constrained non-coding DNA sequences have little or no obvious function (King et al., 2007). This poses two different questions: in the first place, how good is sequence conservation as a guide for function? And in the second place, how many regulatory elements are we missing by only looking at evolutionarily conserved sequences?

Gene deserts are particularly enriched in constrained non-coding sequences in mammals, and in fact a possible explanation for such gene poor regions is that they need to accommodate large number of regulatory elements for genes with complex regulation, such as trans-dev genes (Ovcharenko et al., 2005). However, a recent report challenged this view when two gene deserts, containing multiple regions conserved in mammals, were deleted from the mouse genome with no overt phenotypical effect (Nobrega et al., 2004). These results would imply that sequence conservation is a poor guide for function. Similarly, the pilot results from the ENCODE project could not find any evidence of function for 40% of constrained regions, and that there was no conservation for about 50% of the functional elements identified (King et al., 2007). There is no reason to expect that all cis-elements will be under the same level of constraint, and certainly many genes show differences in expression between human and mouse, and the sequences of cis-elements should have changed in these cases (Valverde-Garduno et al., 2004). It is possible then that a substantial fraction of the regulatory regions in humans (or any species) have been active only recently on an evolutionary time scale (King et al., 2007). In any case, these changes would preferentially relate to novel functions that appeared in a specific lineage.

In an extreme interpretation, these observations would tell us
that evolutionary conservation is irrelevant to function, as we will have a 50% chance that a randomly picked DNA fragment will be functional. However, in both cases described above, only mammals were used to define conserved sequences. These results can therefore be caused by too little phylogenetic depth in the comparison to allow identifying functional elements. Besides, the consequences of gene desert deletion in mice can be difficult to find, and a viable mouse does not mean that no phenotype is present in these strains (Nobrega et al., 2004). More recently, similar results have been obtained, in this case where the deletion of ultraconserved elements does not lead to any phenotype (Ahituv et al., 2007). However, without a complete knowledge of the regulatory landscape of a given gene, the existence of redundant regulatory elements elsewhere that would compensate the deletion cannot be ruled out. With regards to the ENCODE analysis, certainly not all possible assays for function have been performed, so it is difficult to rule out any function at all for 40% of conserved regions (King et al., 2007). In stark contrast, when the phylogenetic filter applied is that of conservation between mammals and fishes, 45% of elements tested in a single assay (mouse transgenics) at a single embryonic stage showed positive activity as cis-regulatory elements (Pennacchio et al., 2006). Similarly, in the survey of the lxB cluster gene deserts, 80% of elements conserved between mouse and pufferfish had enhancer activity in vivo (de la Calle-Mustienes et al., 2005). In conclusion, a careful selection of the range of species used in genomic comparison, or even just of thresholds when comparing closely related species (Prabhakar et al., 2006), is fundamental to reliably identify regulatory elements in the genome.

While the answer to the first question is that conserved sequences are highly likely to be regulatory elements, the second question asks if there can be regulatory elements that have a conserved function but cannot be identified by genomic comparison, or even just of thresholds when comparing closely related species (Prabhakar et al., 2006), is fundamental to reliably identify regulatory elements in the genome.

An illustrating example in this sense comes from the analysis of the regulation of Hox group 3 paralogs in the vertebrate hindbrain (Fig. 3). Mouse Hoxa3 and Hoxb3 are both direct targets of kreisler (MafB), that is expressed and regulates the specification of rhombomeres (r) 5 and 6 (Cordes and Barsh, 1994). cis-regulatory elements that contain functional kreisler binding sites have been identified for both genes in mouse and chick, but while the Hoxa3 element gives a direct read-out of kreisler expression in r5 and r6, the Hoxb3 element is restricted to r5 by the combined action of kreisler and other factors (Manzanares et al., 1997). Zebrafish hoxb3a is expressed in both r5 and r6, resembling Hoxa3 (Hadrys et al., 2006). This suggests that early in vertebrate evolution, both Hoxa3 and Hoxb3 were expressed in r5 and r6 in direct response to kreisler and that the mouse regulatory elements have a common origin and later diverged, resulting in the restriction of Hoxb3 to r5.

When the sequence of the functionally identified Hoxa3 r5/r6 element from mouse and chick is compared, we find a high degree

Fig. 3. Different degrees of sequence similarity in cis-regulatory elements from paralogous genes. (A) Kreisler-responsive regulatory elements from chick and mouse Hoxa3 and Hoxb3 that drive reporter expression in r5–r6 and only r5 respectively, have been functionally characterized (middle and right), kreisler itself is expressed in r5 and r6 (left), but its action on the Hoxb3 element is restricted by other factors. However, while the Hoxa3 element shows high overall similarity between mouse and chick, in the case of the Hoxb3 element, this is restricted to the functional sites (B). This last regulatory element, although conserved in function and in the sequence of critical transcription factor binding sites, is invisible to current comparative genomic approaches as it falls below the normally used thresholds. Functionally tested binding sites for kreisler and other factors are boxed in red, dashes represent gaps, and black boxes and dots identical residues.

...
of conservation (Fig. 3B) that extends up to cartilaginous fishes (Manzanares et al., 2001), and in fact such element has been independently identified in studies of evolutionarily conserved non-coding sequences in vertebrate Hox clusters (Santini et al., 2003). In contrast, the Hoxb3 element does not turn up in any study using current comparative genomic tools. Nevertheless, the alignment of mouse and chick sequences reveals two short stretches of similarity that contain the functional kreasler binding sites (Fig. 3B), that have undoubtedly been constrained by their functional role (Manzanares et al., 1997). This example tells us that even very short regions of sequence conservation may be enough for a conserved function, and that comparative genomics and phylogenetic footprinting are a good indicator of function, but that the need for experimental assays of regulatory activity is undeniable.

**Integration of multiple techniques to unravel cis regulation**

What is becoming clear is that no individual approach will be solely sufficient to fully understand how gene regulation takes place, and integration of data obtained with different methods will be a must. A revealing example of the value of the integration of data obtained from different experimental approaches is shown in Figure 4. Six3 and Six2 are two homebox encoding genes that are close together in a genomic region syntenic in all vertebrates. Both genes show complex restricted expression patterns. Comparative genomic indicate the presence of many highly conserved non-coding regions in this genomic interval (Fig. 4B). It is very likely that different cis-regulatory elements required for different aspects of Six3 and Six2 expression patterns are associated with these conserved regions. Surprisingly, despite their close genomic association and in contrast to other example discussed above, the expression patterns of Six3 and Six2 are completely different. Thus, Six3 is expressed in the forebrain while Six2 is detected in the placodes and in mesodermal derivatives such as the kidneys and muscles. Why do these clustered genes not share expression domains? This dilemma may be partially answered by the in vivo distribution of CTCF in this genomic region (Barski et al., 2007). Using the Chip-Seq technique, it can be shown that the Six3/Six2 intergenic region contains two highly occupied CTCF sites (Fig. 4A). These data suggest that a strong insulator prevents the influence of Six3 regulatory regions on Six2 and vice versa. Moreover, this insulator could well be responsible for keeping both genes associated. Thus, if a breakpoint occurs between Six3 and Six2, the cis-regulatory rich genomic region that does not stay associated with the insulator could have a strong regulatory negative impact in its new genomic neighbourhood. The prediction will be that if the Six3/Six2 intergenic region is deleted in mice, the expression patterns of both genes will become largely overlapping.

A similar situation is found at the Six6/Six1/Six4 genomic locus. These three genes are also clustered, their genomic organization is conserved in all vertebrates, and the genomic region contains several highly conserved non-coding regions that solely sufficient to fully understand how gene regulation takes place, and integration of data obtained with different methods will be a must. A revealing example of the value of the integration of data obtained from different experimental approaches is shown in Figure 4. Six3 and Six2 are two homebox encoding genes that are close together in a genomic region syntenic in all vertebrates. Both genes show complex restricted expression patterns. Comparative genomic indicate the presence of many highly conserved non-coding regions in this genomic interval (Fig. 4B). It is very likely that different cis-regulatory elements required for different aspects of Six3 and Six2 expression patterns are associated with these conserved regions. Surprisingly, despite their close genomic association and in contrast to other example discussed above, the expression patterns of Six3 and Six2 are completely different. Thus, Six3 is expressed in the forebrain while Six2 is detected in the placodes and in mesodermal derivatives such as the kidneys and muscles. Why do these clustered genes not share expression domains? This dilemma may be partially answered by the in vivo distribution of CTCF in this genomic region (Barski et al., 2007). Using the Chip-Seq technique, it can be shown that the Six3/Six2 intergenic region contains two highly occupied CTCF sites (Fig. 4A). These data suggest that a strong insulator prevents the influence of Six3 regulatory regions on Six2 and vice versa. Moreover, this insulator could well be responsible for keeping both genes associated. Thus, if a breakpoint occurs between Six3 and Six2, the cis-regulatory rich genomic region that does not stay associated with the insulator could have a strong regulatory negative impact in its new genomic neighbourhood. The prediction will be that if the Six3/Six2 intergenic region is deleted in mice, the expression patterns of both genes will become largely overlapping.

A similar situation is found at the Six6/Six1/Six4 genomic locus. These three genes are also clustered, their genomic organization is conserved in all vertebrates, and the genomic region contains several highly conserved non-coding regions that

---

**Fig. 4. Combined phylogenetic footprinting and CTCF binding distribution at the Six3/Six2 locus.** The diagrams shows 800 kb encompassing Six3, Six2 and adjacent genes. (A) The distribution of CTCF, determined by Chip-Seq, indicates a high occupancy of CTCF of two sites between Six3 and Six2 genes. (B) Many highly conserved non-coding elements (red bars in intergenic regions and pink bars in introns) are located throughout this region. The exons of the genes are shown in blue.
are likely to contain cis-regulatory region that operate on these genes. As Six3, Six6 is expressed in the anterior neuroectoderm, while Six1 shows an expression pattern similar to that of Six2. The third member of the cluster, Six4, is expressed in a pattern more related to Six1, since it is expressed in the placodes and in mesodermal derived tissues. Again, the Chip-Seq technique indicates a highly occupied CTCF site between Six6 and Six1 but no CTCF accumulation between Six1 and Six4. This indicates that the regulatory regions are, again, divided in two domains separated by an insulator, one regulatory domain acting on Six6 and another on Six1 and Six4. Again, this prediction will require loss of function studies in mice to be verified.

**Regulatory codes: can we find them?**

A regulatory code would be a sequence-based descriptor used to examine primary DNA sequence and identify regulatory elements. Such descriptors would be specific for particular embryonic domains, organs or physiological conditions, and should allow identifying in silico the batteries of genes likely to respond in such domains or conditions. Unfortunately, we still do not have these codes, but it seems extremely possible that they do exist. With them in hand, we could read the regulatory genome as we now read an mRNA sequence to produce a peptide. The problem then is reduced to how to find them.

Major advances in this search have been made in *Drosophila* (Erives and Levine, 2004, Markstein *et al.*, 2004), building on the extraordinary amount of experimental information available on the function of certain classes of transcription factors. The knowledge gained from the detailed study of a small number of regulatory elements responsive to a given transcription factor is used to build a model that describes them, which in turn is used to interrogate the whole genome to find previously undescribed regulatory elements and target genes. These descriptors are heavily based on the clustering of binding sites for more than one factor in a specific regulatory element, a situation usually found in early developmental genes in the fly.

In contrast to this deductive approach, where a clear hypothesis is formulated on what a regulatory element must be like, some recent studies have taken an inductive approach, where the experimental evidence is used to create a model with no prior knowledge of which sites and factors must be acting on that element. The combination of large-scale expression analysis, together with binding site prediction and evolutionary conservation, allows predicting the tissue-specific activity of putative cis-regulatory elements on a whole genome scale (Pennacchio *et al.*, 2007). In an even simpler approach, multiple different regulatory elements driving expression in a specific embryonic structure (the forebrain) were identified in a large-scale transgenic screen of evolutionarily conserved non-coding sequences. These elements were then compared to build a cis-regulatory signature for this structure and test its predictive value (Pennacchio *et al.*, 2006). Other combinations of computational approaches using transcription factor binding affinities have also shown their value for the prediction of regulatory elements (Hallikas *et al.*, 2006).

All of these methods, although still not perfect, at least show that regulatory codes can be found. Maybe we will need bolder approaches, as for example testing randomly overlapping fragments that cover the whole genome by transgenesis, as has been performed in a pilot screen in the sea-squirt *Ciona* (Harafufi *et al.*, 2002). Another possibility would be to take advantage of the low cost and speed of tissue culture assays. Libraries of random overlapping genomic fragments linked to a minimal promoter and a reporter gene could be tested for regulatory activity in a large panel of different cell lines representing as many tissues as possible. The readout from these assays could provide a first approximation to identify fragments with some tissue-specificity, which could subsequently be tested in an *in vivo* transgenic assay.

**Future prospects**

No single experimental approach will be sufficient to understanding the regulatory genome. A combination of computational, *in vitro* and transgenic assays will be needed to scan the DNA sequence in order to find regulatory elements and understand their function in the organism (ENCODE Project Consortium, 2007). On the other hand, we will need as much data as possible on global studies of gene expression, chromatin structure and transcription factor binding site occupancy in as many different conditions as we can manage. Using a comparative approach will be necessary to discriminate between different levels of conservation and how they relate to function (Gomez-Skarmeta *et al.*, 2006). It will also allow identifying a minimal set of regulatory interactions in each biological system. Major efforts must be put in providing conclusive evidence for the function of putative cis-elements identified by other means. At present, the most straightforward way to achieve this is to engineer genomic deletions in mouse, where the effect of the removal or the subtle change of some base pairs in an element can be tested in an otherwise normal animal. We have not talked here about Gene Regulatory Networks (Davidson, 2006), but obviously that will be the next step in understanding how the regulatory genome results in organismal complexity.

Also outside of the scope of this review, but of great interest, is the role of the regulatory genome in evolution and disease. Two papers published during the seventies argued, on the basis of indirect evidence, that cis-regulatory elements might have a critical role in evolution. (Britten and Davidson, 1971; King and Wilson, 1975). The modest degree of divergence in protein sequence cannot account for the profound phenotypic differences between the species, and it has been proposed that regulatory mutations must play a role in the process (Wray, 2007). Once more, the comparison of regulatory networks among species will reveal where changes have occurred that can be correlated with the appearance of evolutionary novelties (Carroll *et al.*, 2001). In a similar fashion, mutations and genomic rearrangements affecting regulatory elements and networks leading to human diseases can be identified and give new clues on their etiology. These are exciting times, and in the near future our understanding of the genome’s second code will unveil novel paradigms and new questions.

**Acknowledgements**

We wish to thanks Susana Cañón and Beatriz Fernandez-Tresguerres for comments and discussions, and Gemma Iglesias for help in preparing Figure 3. Work in our labs was financed by grants from the Spanish Ministry of Education and Science (BFU2005-00025 to MM and BFU2004-
References


Further Related Reading, published previously in the *Int. J. Dev. Biol.*


**TTF-1/NKX2.1 up-regulates the *in vivo* transcription of nestin**
Roberta Pelizzoli, Carlo Tacchetti, Paola Luzzi, Antonella Strangio, Grazia Bellese, Emanuela Zappia and Stefania Guazzi

**Regulation of the mouse alfaB-crystallin and MKBP/HspB2 promoter activities by shared and gene specific intergenic elements: the importance of context dependency**
Shivalingappa K. Swamynathan and Joram Piatigorsky

**Expression and comparative genomics of two serum response factor genes in zebrafish**
Jody L. Davis, Xiaochun Long, Mary A. Georger, Ian C. Scott, Adam Rich and Joseph M. Miano

**Transcriptional regulation and the evolution of development.**
Gregory A Wray

**Regulation of the mouse alfaB-crystallin and MKBP/HspB2 promoter activities by shared and gene specific intergenic elements: the importance of context dependency**
Shivalingappa K. Swamynathan and Joram Piatigorsky

**Interplay of Pax6 and SOX2 in lens development as a paradigm of genetic switch mechanisms for cell differentiation**
Hisato Kondoh, Masanori Uchikawa and Yusuke Kamachi

**Additional enhancer copies, with intact cdx binding sites, anteriorize *Hoxa-7/-lacZ* expression in mouse embryos: evidence in keeping with an instructional cdx gradient**
Stephen J. Gaunt, Adam Cockley and Deborah Drage

**Evolution of cis-regulation of the proneural genes.**
Jean-Michel Gibert and Pat Simpson

**A 3’ remote control region is a candidate to modulate *Hoxb-8* expression boundaries.**
I Valarché, W de Graaff and J Deschamps
Int. J. Dev. Biol. (1997) 41: 705-714

**Analysis of transcriptional regulatory regions in vivo.**
R J MacDonald and G H Swift

**Control of the expression of the Mrf4 and Myf5 genes: a BAC transgenic approach**
JJ Carvajal, D Cox, D Summerbell, PWJ Rigby
Int. J. Dev. Biol. (2001) 45: S139-S140