Deep conservation of cis-regulatory elements in metazoans

Ignacio Maeso1,†, Manuel Irimia2,†, Juan J. Tena3, Fernando Casares3 and José Luis Gómez-Skarmeta3

1Department of Zoology, University of Oxford, Oxford, UK
2The Donnelly Centre, University of Toronto, Toronto, Ontario, Canada
3Centro Andaluz de Biología del Desarrollo (CABD), Consejo Superior de Investigaciones Científicas/Universidad Pablo de Olavide, Sevilla, Spain

Despite the vast morphological variation observed across phyla, animals share multiple basic developmental processes orchestrated by a common ancestral gene toolkit. These genes interact with each other building complex gene regulatory networks (GRNs), which are encoded in the genome by cis-regulatory elements (CREs) that serve as computational units of the network. Although GRN subcircuits involved in ancient developmental processes are expected to be at least partially conserved, identification of CREs that are conserved across phyla has remained elusive. Here, we review recent studies that revealed such deeply conserved CREs do exist, discuss the difficulties associated with their identification and describe new approaches that will facilitate this search.

1. Deeply conserved metazoan features

Animals greatly exemplify how enormous morphological diversity can arise through evolution. However, despite their profound variation in forms, there are clear commonalities in how animal body plans are built. The early formation of distinct germ layers [1–4], the establishment of body axes [5–9], the differentiation of photoreceptors [10] or the deployment of similar neuronal architectures [11] are examples attesting to deep homologies at the tissue and cellular level. Accordingly, whole genome sequencing projects and gene family evolution studies revealed that all animals share a basic set of genes [12–16], establishing a link between this basic shared genetic toolkit and the core of common structures, cell types and developmental processes. Studies focusing on the structure and evolution of developmental gene regulatory networks (GRNs) have further suggested that regulatory gene interactions can also be extremely stable through long periods of evolutionary history [17], which is mirrored by a large collection of deeply conserved gene expression patterns during embryonic development of many animal phyla. These range from the paradigmatic anterior–posterior Hox patterning [18] to commonalities on nervous system regionalization [19–23] and cell-type-specific gene expression patterns [11]. However, in stark contrast to the deep conservation of the protein-coding sequences and expression patterns of the genes involved in common developmental programmes, comparisons of whole genome sequences across metazoans showed that the cis-regulatory elements (CREs) responsible for the expression of these genes do not seem to be conserved among phyla. With a few recently reported exceptions (see below), all highly conserved regulatory modules found to date seem strictly circumscribed to specific taxonomic groups within phyla [24]. Nonetheless, confident identification of conserved CREs is much more complex than that of protein-coding regions, since the structure of the regulatory regions itself—normally consisting of a set of semi-independent short transcription factor binding sites (TFBSs) that can be located at almost unlimited distances from their target genes—may preclude their identification at deep evolutionary distances.

Here, we review studies searching for deeply conserved CREs and discuss whether their results (i.e. widespread lack of deep conservation of CREs) are...
due to true evolutionary peculiarities of CREs, experimen-
tal limitations or a mixture of both. Then, we will
discuss how new experimental approaches based on high-
throughput sequencing combined with data on conserved
positional information will be likely to provide new insights
into these questions.

2. Deeply conserved cis-regulatory elements: beyond the limits of phylogenetic footprinting?

With the advent of whole genome sequencing projects for
many vertebrate species, proper genome-wide comparisons
and phylogenetic footprinting of non-coding regions became
readily available (table 1). These comparisons revealed the
presence of thousands of highly conserved non-coding regions
(HCNRs) among vertebrate genomes, sometimes with levels of
sequence conservation higher than those of protein-coding
regions [25,26,37–39]. The vast majority of these HCNRs
flank genes that play key roles in embryonic development
and have very complex expression patterns [26,39]. Impor-
tantly, functional studies using transgenic assays in mouse,
Xenopus and zebrafish carried out by various groups, includ-
ing ours, showed that a large fraction of the analysed
HCNRs act as transcriptional enhancers, driving expression
of their flanking developmental genes to spatio-temporal
domains consistent with their endogenous expression patterns
[25,26,40–53]. Moreover, similar patterns of HCNRs around
different developmental regulations were subsequently observed
for flies and nematodes [34,35,54–56], and their frequent role
in regulating transcription could also be established.

After the success of the initial studies, identification of
HCNRs thus became the gold standard for the search for func-
tional conserved CREs across model organisms in the past
decade. Within mammals, around 30 000 HCNRs were identified
[57,58], and over 8000 HCNRs were traced back to the origin of
gnathostome vertebrates [27,59]. A set of approximately 2000–
5000 of these HCNRs are currently present in most jawed
vertebrates [27] and were thus suggested to contain much of
the regulatory information essential to build the conserved
vertebrate body plan [24,28], providing a unifying view of
conserved embryonic development and body plan, gene toolkit
and gene expression and its underlying regulatory circuitry.

However, a major challenge to this view came from the
sequencing of the slow-evolving invertebrate chordate
amphioxus. Compared to over 5000 HCNRs present in gnatho-
stome ancestors, only around 50 non-coding regions showed
recognizable homology by sequence similarity between the
amphioxus and human genome [32], despite extensive body
plan and developmental similarities within the chordate
phylum [60,61]. Recently, we performed a wide survey for
conservation of these HCNRs in multiple metazoan lineages
using local alignments of these sequences to conserved synten-
ic regions combined with classic phylogenetic footprinting.
We identified eight deeply conserved HCNRs, two of them
already present in the eumetazoan ancestors, and six of
them in the last common ancestor of deuterostomes [36]. Func-
tional studies of these sequences in zebrafish, sea urchin and
Drosophila revealed that they promote expression in similar tis-
sues (figure 1), as would be expected for homologous CREs
read by deeply conserved regulatory states [36]. In a similar
study, Clarke et al. [30] extended the conservation of two of
the deuterostome-specific HCNRs to the origin of bilaterians
and showed their ability to drive expression in different
animal systems. Altogether, these analyses suggest that identi-
fication of deep HCNRs is largely beyond the range of
traditional phylogenetic footprinting analyses. The reported
sequences, although clearly homologous, showed a reduced
degree of conservation than that observed within ver-
tebrates—usually a core of only 75–100 bp with 60–70% of
similarity compared with often greater than 500 bp within
vertebrates—making them very difficult to be computationally
detected, especially if we take into account that it is likely that
in many other cases transphylytic CREs would not be config-
ured as HCNRs (table 1). Nevertheless, these initial studies
show that ancient CREs conserved across metazoan phyla can
and do exist, raising the question of whether they may represent
rare exceptions or only the tip of the iceberg emerging from a
sea of technical and perhaps also conceptual limitations.

3. Challenges of in silico detection of transphylytic cis-regulatory elements based on sequence similarity

The shorter length and lower similarity of the HCNRs con-
taining the few described transphylytic CREs place them on the
edge of the cut-offs that have been traditionally used to
assess homology based on sequence alignment scores [62].
Therefore, this approach is likely to lead to a much higher
rate of false negatives when looking for conservation of
CREs between extremely divergent lineages, amplifying the
effects of common confounding factors of phylogenetic
footprinting, such as species selection [62]. For instance, in the
recently discovered bilaterian CREs in protostome genomes,
the sequences from amphioxus and zebrafish were the only
deuterostome queries that were able to recover their proto-
tostome orthologues. In fact, in all cases but two, amphioxus
transphylytic CREs displayed the highest degree of similarity
with their orthologues in other phyla, highlighting the impor-
tance of the inclusion of slow-evolving taxa [30]. Along the
same lines, Taher and collaborators [63] have recently
shown the utility of applying conservation ‘tunnelling’ to
identify divergent orthologous CNRs. This method consists
of a series of iterative pairwise alignments among sequences
from multiple species diverged at different times and under
distinct evolutionary rates. By ‘tunnelling’ human–zebrafish
comparisons through the inclusion of frog sequences, the
authors were able to identify 1500 pairs of sequences with
no clear similarity between human and zebrafish, but that
were clearly orthologous (more than 70% over more than
100 bp) to the frog counterparts.

Unfortunately, these difficulties are probably not solvable
by simply relaxing the cut-offs of length and sequence simi-
larity used to define conserved elements: the rate of false
positives would increase dramatically and, owing to the speci-
fic nature of CREs, other detection artefacts may appear. For
instance, the presence of combinations of relatively long TFBS
(e.g. the 15–20 bp long HMG/POU cassette bound cooperati-
vely by Sox2 and Oct4 proteins [64]) around different
genes sharing a similar regulatory syntax could lead to the
appearance of multiple instances of similar non-homologous
sequence stretches in a given genome. These kinds of
Table 1. Examples of studies searching deeply conserved non-coding regions at different animal evolutionary timescales. In each study, the number of CNRs identified, tested and positive, as well as the percentage similarity and lengths of these CNRs is depicted. n.a., not assessed; n.s., not specified; (a), average; (m) median.

<table>
<thead>
<tr>
<th>Phylogenetic range</th>
<th>Study</th>
<th>Species/lineages compared</th>
<th>No. elements</th>
<th>No. elements positive/tested</th>
<th>Positional conservation</th>
<th>Min. similarity (%)</th>
<th>Ave. similarity (%)</th>
<th>Ave./med. length (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>bony vertebrates</td>
<td>[25]</td>
<td>human, fugu</td>
<td>3124</td>
<td>57/137</td>
<td>yes</td>
<td>70</td>
<td>n.s.</td>
<td>n.s.</td>
</tr>
<tr>
<td></td>
<td>[26]</td>
<td>human, rodents, chicken, fugu</td>
<td>1373</td>
<td>23/25</td>
<td>n.s.</td>
<td>74</td>
<td>84</td>
<td>199 (a)</td>
</tr>
<tr>
<td>jawed vertebrates</td>
<td>[27]</td>
<td>elephant shark, tetrapods, teleosts</td>
<td>8452</td>
<td>307/564</td>
<td>n.s.</td>
<td>74</td>
<td>82</td>
<td>175–226 (a)</td>
</tr>
<tr>
<td>vertebrates</td>
<td>[28]</td>
<td>tetrapods, teleosts, lamprey</td>
<td>73&lt;sup&gt;a&lt;/sup&gt;</td>
<td>2/2</td>
<td>yes&lt;sup&gt;d&lt;/sup&gt;</td>
<td>63</td>
<td>80</td>
<td>127 (a)</td>
</tr>
<tr>
<td>olfactory</td>
<td>[29]</td>
<td>vertebrates, sea squirts</td>
<td>183</td>
<td>2/3</td>
<td>no</td>
<td>52</td>
<td>55</td>
<td>45 (a)</td>
</tr>
<tr>
<td>chordates</td>
<td>[30]</td>
<td>vertebrates, amphioxus</td>
<td>5</td>
<td>2/2&lt;sup&gt;c&lt;/sup&gt;</td>
<td>yes</td>
<td>66</td>
<td>71</td>
<td>108 (a), 112 (m)</td>
</tr>
<tr>
<td></td>
<td>[31]</td>
<td>vertebrates, amphioxus</td>
<td>1299</td>
<td>10/22</td>
<td>partial (gene linkage only)</td>
<td>n.s.</td>
<td>n.s.</td>
<td>54 (m)</td>
</tr>
<tr>
<td></td>
<td>[32,33]</td>
<td>human, amphioxus</td>
<td>56</td>
<td>4/8</td>
<td>yes</td>
<td>60</td>
<td>64</td>
<td>63 (m)</td>
</tr>
<tr>
<td>fruitflies</td>
<td>[34]</td>
<td>D. melanogaster, D. anassae, D. pseudoobscura, D. viridis, D. mojavensis</td>
<td>6779</td>
<td>n.a.</td>
<td>yes</td>
<td>98</td>
<td>n.s.</td>
<td>59 (m)</td>
</tr>
<tr>
<td>Cenarchobditis</td>
<td>[35]</td>
<td>C. elegans, C. briggsae, C. remanei</td>
<td>2084</td>
<td>12/n.s.&lt;sup&gt;b&lt;/sup&gt;</td>
<td>n.s.</td>
<td>100 in 30 bp</td>
<td>96</td>
<td>69 (a), 59 (m)</td>
</tr>
<tr>
<td>nematodes</td>
<td>[30]</td>
<td>vertebrates, amphioxus, sea urchin, acorn worm</td>
<td>4</td>
<td>2/2&lt;sup&gt;c&lt;/sup&gt;</td>
<td>yes</td>
<td>59</td>
<td>65</td>
<td>104 (a), 108 (m)</td>
</tr>
<tr>
<td></td>
<td>[36]</td>
<td>vertebrates, amphioxus, sea urchin, acorn worm</td>
<td>8</td>
<td>2/8</td>
<td>yes</td>
<td>55</td>
<td>65</td>
<td>103 (a), 119 (m)</td>
</tr>
<tr>
<td>Bilateria</td>
<td>[30]</td>
<td>vertebrates, amphioxus, sea urchin, acorn worm, tick, gastropods</td>
<td>2</td>
<td>2/2&lt;sup&gt;c&lt;/sup&gt;</td>
<td>yes</td>
<td>55</td>
<td>61</td>
<td>99 (a,m)</td>
</tr>
<tr>
<td>Eumetazoans</td>
<td>[36]</td>
<td>vertebrates, amphioxus, sea urchin, acorn worm, anemone</td>
<td>2</td>
<td>1/2</td>
<td>yes</td>
<td>57</td>
<td>63</td>
<td>117 (a,m)</td>
</tr>
</tbody>
</table>

<sup>a</sup>Probably an underestimation, only a subset of gnathostome CNRs were searched in the incomplete lamprey genome assembly.

<sup>b</sup>Data obtained by overlap with previously published results.

<sup>c</sup>The elements are the same.

<sup>d</sup>Owing to incompleteness of the lamprey assembly only one syntenic block could be checked.
Figure 1. A deeply conserved HCNR located downstream of the Sox2 gene from several species promotes GFP expression in similar neural territories in stable zebrafish transgenic lines.

phenomena could indeed provide an explanation for the striking differences in the complements of chordate HCNRs reported by several studies. In stark contrast with the scarcity of ancient HCNRs identified by us and others [30,32,36], two reports using the same methodology [65] uncovered 1299 and 183 putative HCNRs dating back to the last common ancestor of chordates (i.e. conserved between amphioxus and vertebrates [31]) and of the olfactores lineage (the clade comprising tunicates plus vertebrates [29]), respectively. The majority of these sequences were extremely short, with an average of 45 bp in the olfactores sequences and a median of 54 bp in the case of the chordate regions. Similarly, the percentage identities were also very modest (55% on average between tunicates and vertebrates). Finally, the relative location of these non-coding sequences was only partially conserved between vertebrates and invertebrates. In the case of the amphioxus–vertebrates study, although the putative HCNRs were linked to the same orthologous genes in both chordate lineages, the conservation of their relative position and orientation with respect to these genes (i.e. intronic, intergenic upstream/downstream or within a nearby bystander gene) was not taken into account, thus weakening the arguments supporting the homology of the amphioxus and vertebrate regions. For the putative olfactores HCNRs, the lack of positional conservation was particularly extreme: in 182 of the 183 reported cases it was not possible to identify a single orthologous gene that was syntenic to the non-coding sequences in both vertebrates and tunicates, even when using 2 Mb windows, which span an average of 351 genes in Ciona intestinalis [66] (in the remaining case, the HCNR was located within an intron of FoxP1 in both vertebrates and tunicates). Thus, only one of the putative olfactores HCNRs could have the same target in vertebrates and invertebrates, casting doubt on the homology (and hence conservation) of these sequences. In the absence of positional conservation, establishing homology based on the presence of very short stretches of moderate sequence similarity seems, at best, risky. Therefore, although some of the sequences reported by Huf ton et al. [31] and Sanges et al. [29] could correspond to bona-fide conserved homologous non-coding sequences, it is possible that the number of HCNRs reported in these studies is an overestimation.

In summary, using current computational tools, the number of unambiguous transphyletic HCNRs identified to date is surprisingly small. Indeed, given the challenges faced by HCNRs search, the question may no longer be why we are not able to detect transphyletic CREs using sequence conservation but, rather, why we can actually identify (at least some of) them as HCNRs.

4. Why so few transphyletic cis-regulatory elements?

The difficulties described above raise the question of whether the nearly complete lack of known CREs conserved across phyla represents true evolutionary peculiarities of cis-regulatory modules compared to protein-coding sequences, it responds to the technical and conceptual challenges associated with CRE identification and/or comparison based on sequence conservation, or something in between. Therefore, at least three major evolutionary frameworks can be proposed for this differential pattern of conservation:

(i) The lack of conservation of CREs may parallel the specific differences in body plans across phyla [24]. In this scenario, a conserved set of protein-coding sequences would have been recurrently redeployed for different developmental programmes by evolving new CREs in the different lineages; therefore, the conservation of CREs would be minimal across phyla and would be represented by the few transphyletic HCNRs identified to date. Although this may be the case for a large fraction of phylum-specific CREs, it does not provide a satisfactory explanation for the CREs associated with the large number of deeply conserved expression patterns and developmental programmes described above.

(ii) Rampant turnover of CREs may have fully eroded the conservation of the ancestral set of CREs, while keeping
the ancestral functions [67–69]. If so, an analogous, but no longer homologous, set of CREs would be responsible for the similar gene expression patterns and functions in different phyla. This idea has been boosted by a series of recent genome-wide studies of CRE conservation that showed a very high TFBS turnover, even between relatively closely related species [69,70]. Behind this rapid rate of TFBS gain/loss is probably the high degree of (partial) functional redundancy observed for many regulatory inputs, by which several CREs often drive expression of the same gene to similar spatio-temporal locations [42,43,71–73]. However, although this rampant turnover may be widespread for simple TFBSs, particularly in adult tissues, it may not be so extreme for more complex CREs associated with genes involved in regulation of embryonic development [74], especially those CREs forming intricate enhanceromes (i.e. long arrays of TFBSs that work cooperatively and behave as functional and evolutionary blocks, often resulting in HCNRs) [75,76]. Furthermore, it is still a matter of debate how many of the whole set of TFBSs identified by ChIP-based studies are actually biologically functional (see [77]) and to what extent they contribute to the transcriptional regulation of their surrounding genes [78].

(iii) Deeply conserved ancestral CREs do exist in various phyla, but, as discussed above, we cannot recognize their orthology by mere sequence similarity in alignment procedures, either because they are not configured as HCNRs and have diverged too much, or because of technical limitations of current computational tools and/or an insufficient and biased taxon sampling. In this scenario, conserved ancestral CREs would have evolved from a common set of ancestral instructions that have diverged in sequence beyond detection level, but preserved their functional properties to a large extent. Therefore, while in (ii) a newly created CRE takes over the ancestral function and the ancestral CRE disappears, in this case, although mutations can erase the ancestral sequence similarity, these mutations occur within the same orthologous CRE that has maintained its functionality uninterruptedly over evolutionary time. Suggesting that this process may be widespread, hundreds of orthologous ‘covert CREs’ (i.e. elements with conserved regulatory components and syntax embedded within divergent sequences that drive expression to similar anatomical domains) have been identified between human and zebrafish in conserved syntenic positions [63].

Under this last framework, however, two major issues will be encountered when trying to define the complement of ancestral transphyletic CREs. First, the methods used for detecting CREs have to be fully independent of sequence similarity (i.e. phylogenetic footprinting), and based on specific functional and biochemical properties; fortunately, approaches based on next-generation sequencing (NGS) can currently provide unprecedented power for CRE detection. Second, assessing whether two genes inherited ancestral CREs or they independently evolved into functionally equivalent CREs relies on a large extent on positional information and genome structure, which needs to be taken into consideration, in combination with more sophisticated sequence comparison methods based on conserved regulatory encryption [63].

5. How to look for cis-regulatory elements regardless of sequence information: the epigenomics revolution

Applications of NGS have proved to be the most efficient methods for genome-wide identification of CREs [79–97]. These studies, performed in various model systems, have shown the association between a precise chromatin environment and transcriptional enhancers (i.e. CREs that promote transcription, in contrast to silencers or other types of CREs). Enhancers that correlate with enrichment of lysine 4 mono-methylation of histone H3 (H3K4me1) and lysine 27 acetylation of histone H3 (H3K27ac) modifications are located within DNase hypersensitive sites and are able to recruit transcriptional activators, such as p300 and CBP [80–83,85–89]. Moreover, enhancers can be subdivided into distinct functional classes based on their epigenetic makeup: whereas H3K4me1 alone marks poised enhancers, active ones are enriched in both H3K4me1 and H3K27ac modifications [83,86–89]. Importantly, since these combinations of histone marks have been repeatedly found in different bilaterian species, ranging from flies to mammals, they probably represent evolutionarily conserved signatures of enhancers. Moreover, the coding sequences of histones are extremely conserved, which allows the use of commercially available ChIP-grade antibodies in almost any animal species, facilitating genome-wide mapping of enhancers across many lineages and biological systems.

For instance, NGS-related approaches have been used to determine the dynamics of enhancer activation during cell differentiation, both using embryonic stem cell differentiation strategies in cell cultures experiments [86,88,95–99] or developing embryos [83,89]. In one of these studies, we have shown that deposition and removal of H3K27ac is very dynamic during early zebrafish development and correlates well with the expression levels of the putative target genes of these marked enhancers [83]. Genomic regions enriched in H3K27ac at early stages are bound by pluripotent factors and other control genes involved in early developmental processes. These marks are removed later from these regions and substituted by others in regions that will be bound by tissue-specific transcription factors and/or controlling the expression of patterning genes.

In addition, we found that the putative enhancers that are turned on at late gastrula have the highest sequence conservation within vertebrates, compared with those activated at other developmental stages. Interestingly, the segmentation stages that follow gastrulation show the highest degree of transcriptome conservation across vertebrate development, supporting the ‘hourglass’ model of developmental evolution, which states that maximum morphological similarity between vertebrates occurs at the phylotypic stage [100–103]. Our results thus suggest that such a conserved transcriptionome at the phylotypic stage is controlled by the most evolutionarily conserved set of CREs. Moreover, this high sequence and, presumably, functional conservation makes late gastrula stage enhancers arguably the best candidates to search for functionally equivalent CREs in other lineages.

Another powerful NGS-based method is DNAse I footprint, which allows the identification of transcription factor
occupancy sites at the nucleotide resolution and without the use of specific antibodies [93,104,108]. Strikingly, the sequence conservation of the occupied binding sites is much higher than that detected across the much broader ChIP-seq peaks that contain them [93]. This suggests that even within largely non-conserved regulatory regions, specific conserved binding sites for particular transcription factors may be identified. Moreover, this technique has been used to identify occupied TF binding sites within CREs in a variety of cell lines, allowing the inference of cell-type-specific GRNs [106]. Since specific antibodies are not required for this method, it is likely that DNase I footprinting in multiple animal models will be a particularly powerful strategy to identify evolutionarily conserved GRNs operating along different development stages even in distantly related species.

The combination of all these epigenomic strategies will help to precisely define the collection of CREs active in a particular cell type, tissue or embryonic stage of a given species. Nevertheless, to infer which genes are actually regulated by these CREs is not trivial, and needs to be assessed by different approaches. The extent of DNA containing all regulatory elements acting on a particular gene has been denominated the regulatory landscape of such gene [47,107]. These landscapes can be very large, especially for developmental genes, and composed of multiple regulatory elements scattered across several megabase pairs [108]. Various genetic techniques, including mouse genetic engineering, BAC (Bacterial Artificial Chromosome) recombiner coupled to transgenesis and Chromosome Conformation Capture (3C), have been successfully used to define regulatory landscapes of different developmental genes (e.g. [42,47,109–112]). Recently, a new variant of the 3C technique coupled to NGS (4C-seq) has been shown to be very effective in defining gene regulatory landscapes [113,114]. Indeed, the combination of 4C-seq with epigenomic data, such as those described above, has been used to identify not only tissue-specific Hox CREs, but also the topological organization of Hox regulatory landscapes along the anterior–posterior axis of mouse embryos [113,114] and along limb development [115]. Therefore, the integration of multiple NGS techniques is likely to facilitate the identification of potential evolutionary equivalent CREs in different model organisms.

6. Where to look for functionally conserved cis-regulatory elements: positional information

With the lowering of sequencing costs, the universal application of these NGS-based methods is expected to soon yield a wealth of putative CREs in a wide range of metazoan lineages. The evolutionary challenge, then, will be to assess the putative conservation of functionally equivalent CREs (i.e. driving expression to similar spatio-temporal domains) associated with orthologous genes. In addition to insights provided from improving our understanding of TFBS syntax and its evolution, an obvious reference should be conserved positional, or syntenic, information. Within lineages, HCNRs are located at conserved syntenic positions in different genomes [35,38,39]; this is also true for confidently identified transphyletic HCNRs [30,36]. Therefore, it is expected that most, if not all, ancestral CREs will be found in similar relative positions in divergent animal species. favouring this hypothesis, functionally equivalent enhancers that are not conserved at the sequence level have been found at similar positions relative to their target genes in distantly related insects, supporting their common origin [116]. Similarly, ‘covert CREs’ with no sequence similarity but conserved arrangement of putative TFBSs have been identified in the same relative syntenic positions in human and zebrafish, using coding genes and HCNRs as syntenic anchors [63].

The localization of CREs far apart from the genes they regulate, particularly in the introns of nearby genes (known as bystanders), is known to influence the conservation of gene synteny within phyla, establishing genomic regulatory blocks (GRBs) around developmental genes [34,117]. We have recently shown that this is true even across phyla. A genome-wide survey of 17 different species revealed nearly 600 pairs of unrelated genes that have remained tightly physically linked in diverse lineages over hundreds of millions of years of evolution [118–120]. Importantly, a significant fraction of these conserved associations may correspond to ancient GRBs [119]. Therefore, it is plausible that some of the introns of the conserved bystander genes contain deeply conserved CREs, despite sequence similarity having been eroded. Some preliminary data from our laboratory support this possibility (figure 2). Combining enhancer-associated epigenetic marks in zebrafish embryos, enhancer transgenic assays and 3C experiments, we have shown that an otx1 brain enhancer element lies within one specific intron of the nearby bystander ehbp1 gene [119]. The expression promoted by this enhancer perfectly matches the anterior expression of otx1 in all vertebrates [121]. This expression domain is also very similar to that observed in amphioxus [122], with very limited non-coding sequence conservation with vertebrates [32,36], but with hundreds of conserved gene associations, including between the Otx and Ehbp1 orthologues [119]. Despite no clear sequence conservation being detected in the orthologous Ehbp1 intron between amphioxus and vertebrates, transient transgenic enhancer assays in zebrafish embryos with the amphioxus orthologous intron sequence reveals an enhancer activity highly similar to that of the zebrafish enhancer (figure 2). Therefore, these results suggest that functionally equivalent CREs can be found in similar syntenic positions, strongly arguing for their homology. If so, the finding of hundreds of deeply conserved syntenic associations due to cis-regulatory constraints, presumably from shared ancestral CREs, would provide another indication that conservation of ancestral CREs may be much more widespread than currently assumed. Despite being strongly suggestive, it should be noted, however, that, in the absence of sequence similarity or exact conserved TFBS arrangements, whether functionally equivalent enhancers in conserved syntenic positions of two different species are homologous or analogous cannot be fully elucidated with our current knowledge of CRE evolution, imposing a new exciting challenge for evolutionary genome biologists.

7. Concluding remarks

Current data suggest that a few ancestral CREs exist in present-day animal genomes configured as HCNRs. With these data on hand, the long-term evolution of HCNRs seems radically different from that of the gene complement: the number of HCNRs decreases with the phylogenetic distance, but this decrease is not linear, dropping abruptly at certain evolutionary transitions, often coincidental with transphyletic boundaries, in a pattern highly reminiscent of the evolution of microRNAs [28,30].
However, it is likely that the huge majority of conserved transphyletic CREs are much less constrained in sequence than these few described examples, and we may be able to identify them through a combination of epigenetic profiling, genome architecture comparisons and functional studies. Although functional CRE testing is currently a limiting factor (particularly given that putative orthologous CREs should be ideally tested in their multiple endogenous contexts, a very uncommon practice to date [30,36], and for which reliable techniques are not yet established in many systems), it is likely that new methods combining NGS with computational techniques will not only predict CRE activity, but also accurately infer the expression patterns that they drive; the first efforts in this direction are promising [123]. Finding the set of ancestral CREs and their functions opens up the extremely exciting possibility of unravelling GRNs shared across phyla. These networks would define the basic developmental operations common to all metazoans, and the foundation for the further diversification of their body organization.

Acknowledgements. The authors would like to thank Gill Bejerano and Shoa L. Clarke for providing information necessary for the completion of table 1. We would like to apologize to all researchers whose work was not included in this review.


References


3. Martindale MQ, Hejnol A. 2009 A developmental perspective: changes in the position of the


