Development 140, 1385-1395 (2013) doi:10.1242/dev.084459

Conserved non-coding elements and *cis* regulation: actions speak louder than words

Andrew C. Nelson*,‡ and Fiona C. Wardle‡

Summary

It is a truth (almost) universally acknowledged that conserved non-coding genomic sequences function in the *cis* regulation of neighbouring genes. But is this a misconception? The literature is strewn with examples of conserved non-coding sequences being able to drive reporter expression, but the extent to which such sequences are actually used endogenously *in vivo* is only now being rigorously explored using unbiased genome-scale approaches. Here, we review the emerging picture, examining the extent to which conserved non-coding sequences equivalently regulate gene expression in different species, or at different developmental stages, and how genomics approaches are revealing the relationship between sequence conservation and functional use of *cis*-regulatory elements.

Key words: ChIP, Conserved non-coding elements, *Cis* regulation, Enhancer assay, Phylotypic stage

Introduction

During embryonic development, gene expression must be controlled precisely both spatially and temporally. This control is brought about, in large part, by the combinatorial interaction of specific transcription factors (TFs) with *cis*-regulatory modules (CRMs; see Glossary, Box 1), which are usually located in non protein-coding genomic sequence (Davidson, 2006). An understanding of gene regulation is key if we are to understand the mechanisms by which correct expression is achieved during development, and how variations in these mechanisms drive phenotypic change. Identification of CRMs and the study of their functional makeup is therefore fundamental to developmental biology.

It is generally believed, analogous to the high conservation of coding regions relative to surrounding sequence, that functional regulatory elements will be conserved across evolution. This assumption is supported by the known association of conserved non-coding sequences with developmentally important genes, as well as the identification of conserved non-coding sequences that can drive gene expression during development (reviewed by Elgar, 2009; Vavouri and Lehner, 2009). However, recent studies (discussed below) have provided evidence that conserved *cis*-regulatory sequences do not always have conserved function, and functionally conserved CRMs do not always have conserved sequence. What then is the significance of these sequences? We consider this question with reference to recent pan-evolutionary and pan-developmental studies that throw some light on the relationship between conserved non-coding sequence and *cis* regulation during development, and suggest that conserved regulation may be a feature of the phylotypic stage.

Randall Division of Cell and Molecular Biophysics, New Hunt's House, King's College London, Guy's Campus, London SE1 1UL, UK.

*Present address: Sir William Dunn School of Pathology, University of Oxford, South Parks Road, Oxford OX1 3RE, UK
‡Authors for correspondence (andrew.nelson@path.ox.ac.uk; fiona.wardle@kcl.ac.uk)

CRM identification and function

Methods for finding and studying CRMs

In this Review, we categorize methods of CRM prediction as either 'indirect' or 'direct'. Indirect methods use genomic sequence alone to predict CRMs, for example by identifying clusters of known transcription factor binding sites (TFBSs; see Glossary, Box 1) in a single or multiple species. Alternatively – especially where there is no prior knowledge of cooperating TFBSs – multi-species genomic alignment can be used to identify conserved non-coding regions that may act as CRMs, based on the assumption that the sequence is conserved because an essential regulatory function of that sequence is under selection. Direct methods of identifying CRMs rely on techniques such as DNase hypersensitivity (see Glossary, Box 1) and immunoprecipitation of TFs and other chromatin components that are associate with regulatory DNA (ChIP; see Glossary, Box 1).

The most common method of studying CRM function is the enhancer assay, where the predicted CRM is used to drive a reporter gene in an embryo, revealing the spatiotemporal pattern of expression driven by that CRM. Testing a CRM in multiple species can also show the extent to which it drives conserved expression, something that might be expected of a CRM that is located in a conserved sequence. However, as discussed below, there is often a relatively poor correlation between sequence conservation and functional conservation. In view of this, we suggest that, at this time, direct methods represent a better method for identifying regulatory regions across the genome, although it is likely that an increased understanding of the sequence requirements for different CRMs will emerge as more are mapped and characterized this way. In turn, such mapping and characterization will inform indirect methods of detection.

Conserved non-coding sequence, CRMs and developmental genes

Since the release of the first animal genome sequences (*C. elegans* Sequencing Consortium, 1998; Adams et al., 2000; McPherson et al., 2001; Venter et al., 2001; Aparicio et al., 2002) it has become clear that some regions of non-coding DNA have remained highly conserved across millions of years of evolution. Multi-species alignment of genomic sequences has identified elements that are described as conserved non-coding elements (CNEs; see Glossary, Box 1), highly conserved non-coding elements (HCNEs) or ultra-conserved elements (UCEs), depending on the level of conservation (Bejerano et al., 2004; Sandelin et al., 2004; Woolfe et al., 2005) (see Box 2). For simplicity, we refer to all elements in the various studies described below as CNEs.

Both vertebrate and invertebrate genomes contain CNEs, although studies indicate that vertebrate, insect and nematode

Box 1. Glossary

'Active' regulatory regions. Genomic regions that actively regulate their target genes, characterized by active epigenetic marks, depletion of repressive marks and nucleosome depletion. Active promoters are bound by RNA polymerase II.

Chromatin immunoprecipitation (ChIP). A method used to isolate DNA sequences associated with a protein of interest. Briefly, cells are fixed, and chromatin is extracted and fragmented. Protein-DNA complexes are isolated using antibodies. The bound DNA fragments are then identified by PCR, microarray hybridization or sequencing.

***Cis*-regulatory module (CRM).** A genomic region linked to a target gene and influencing its expression. A single CRM can regulate multiple genes and a single gene can be regulated by multiple CRMs. A CRM may be a promoter, enhancer, insulator or silencer. CRMs contain binding sites for transcription factors.

Conserved non-coding element (CNE). A non-coding region of the genome identified by conventional alignment of genomic sequences from two or more species.

CRM grammar. Sequence organization within CRMs, such as consistent spacing, number and orientation of transcription factor-binding sites, that allows their sequence-based (indirect) identification.

DNase hypersensitivity assay. The assessment of the chromatin state of genomic regions through treatment with DNase. 'Open' (nucleosome depleted) chromatin, which represents functionally active sequences, is more susceptible to DNase degradation. Degraded and retained genomic regions can be identified by PCR, Southern blot, microarray hybidization or sequencing.

Enhancer. A genomic region that enhances the transcription of a target gene(s). Enhancers are often in close proximity to their target genes but can also act at a greater distance. Enhancers can be upstream, downstream or internal to their target genes.

Insulators. A genomic boundary element that blocks the interaction between enhancers and promoters, defining the set of genes that can be regulated by an enhancer. Insulator activity is thought to occur through the 3D structure of DNA mediated by CTCF.

Phylogenetic footprinting. A technique to identify potential CRMs within conserved non-coding sequences through comparison with orthologous sequences from one or more other species.

Phylotypic stage. Stage or stages in embryonic development where species within a phylum show less morphological diversity.

'Poised' promoters/enhancers. *Cis*-regulatory elements characterized by the presence of transcription factors, RNA polymerase II occupancy and epigenetic marks, consistent with, but lacking, active transcription.

Promoter. A genomic region required for initiation of transcription of the proximal gene. Promoters bind the basal transcriptional machinery and determine sites of transcription initiation.

Silencer. A genomic regulatory element capable of binding transcriptional repressors, which prevent RNA polymerase from initiating transcription.

***Trans-dev* gene.** A gene functionally associated with transcriptional regulation and/or development. Such genes are more likely to be associated with CNEs than are other genes.

Transcription factor binding sites (TFBSs). Short genomic sequences that are recognized and can be bound by transcription factors. TFBSs may have a strong affinity for multiple transcription factors, usually of the same family, or be specific to individual ones. TFBSs are the functional components of CRMs, in which they occur in clusters.

CNEs are generally not related to each other at the sequence level (Glazov et al., 2005; Siepel et al., 2005; Vavouri et al., 2006). However, a recent study has identified two elements conserved between vertebrates and invertebrates (Clarke et al., 2012), and it is possible that more will be discovered with advances in computational tools. Interestingly, CNEs are enriched in regions around genes that encode transcriptional regulators or factors involved in embryonic development (*trans-dev* genes; see Glossary, Box 1) (Sandelin et al., 2004; Woolfe et al., 2005) (reviewed by Elgar, 2009), leading to the idea that CNEs are CRMs that control gene expression during normal embryonic development. Since this discovery, numerous studies have shown that conserved sequences can act as CRMs and drive gene expression in embryos (e.g. Nobrega et al., 2003; Johnson et al., 2004; Teng et al., 2004; Woolfe et al., 2005; Pennacchio et al., 2006; Li et al., 2010), suggesting that these sequences may be conserved because of a consistent gene regulatory function. These observations are the basis for phylogenetic footprinting (see Glossary, Box 1) as a means to identify functional regulatory sequences. Another concept that has emerged from studies in both vertebrates and invertebrates is that of the gene regulatory block. This is a region of conserved synteny that encompasses a *trans-dev* target gene, associated CNEs that can regulate the target gene (often at long range), together with bystander genes that are functionally unrelated to the target gene and are not under the control of the CNEs (Engström et al., 2007; Kikuta et al., 2007). It is thought that these CNEs remain in synteny owing to their association with the target gene, whereas bystander genes can escape this region over time because they are not regulated by the conserved element. This is seen, for example, in teleost fish, which have undergone an additional round of genome duplication and rediploidization: the target gene and associated conserved regulatory regions are maintained after duplication, but bystander genes are lost. Thus, these analyses can be used to link CNEs to the regulated gene in a genomic block (Kikuta et al., 2007).

Expression regulation by sequence-conserved CRMs

If CNEs act as enhancer CRMs, they should be able to drive gene expression; if they are conserved because of functional restraint, they should drive gene expression in the same domains across different species. These hypotheses can be tested by fusing the CNE to a reporter gene with a minimal promoter and assaying reporter gene expression during development. Although numerous studies have used this approach to show that, indeed, many CNEs drive specific patterns of expression (e.g. Nobrega et al., 2003; Johnson et al., 2004; Teng et al., 2004; Woolfe et al., 2005; Pennacchio et al., 2006; Li et al., 2010), other studies also demonstrate that where conserved sequence drives expression, the expression pattern is not always conserved across multiple species, and lineage-specific changes in *cis* or *trans* regulation have occurred (e.g. Gordon and Ruvinsky, 2012). Here, we highlight a handful of studies that illustrate the complex relationship between sequence and function.

Woolfe and colleagues identified over 1400 CNEs that are conserved between human and fugu, and tested elements clustered around *sox21*, *pax6*, *hlxb9* and *shh* for enhancer activity in zebrafish (Woolfe et al., 2005). In vertebrate development, zebrafish are a popular choice for reporter studies, as transgenesis assays can provide a quick, efficient and relatively inexpensive readout of expression regulation at multiple developmental time points. Of 25 fugu CNEs used to drive a GFP reporter (i.e. fugu sequences tested in zebrafish, or FZ), over 90% acted as enhancers during mid-embryogenesis. These enhancers were able to drive expression in areas associated with the known expression pattern of the nearest gene (e.g. in the nervous system for *sox21*), although

Box 2. Terminology and methods for identifying conservation

Multi-species genomic alignment can be used to identify conserved non-coding regions that may act as *cis*-regulatory modules (CRMs). The many studies that have used sequence alignment apply a range of criteria to identify such regions. These include percentage sequence identity, size of the genomic region and evolutionary distance between species. Although the terminology used in the literature is intended to highlight the extent to which a genomic sequence is conserved, the range of criteria to which these terms are applied may be misleading. The most typically used labels are conserved non-coding elements [CNEs, 60-70% sequence identity over 100 bp (see Dermitzakis and Clark, 2002; Li et al., 2010)]; highly conserved non-coding elements [HCNEs, 70-98% sequence identity over 30-100 bp (see Sandelin et al., 2004; Woolfe et al., 2005; Engström et al., 2007; Kikuta et al., 2007; Engström et al., 2008)]; and ultra-conserved regions/elements [UCEs, 95-100% sequence identity over 200 bp (see Bejerano et al., 2004; Pennacchio et al., 2006)]. Although UCEs appear to exist in the order of hundreds in vertebrate genomes, and HCNEs are of the order of tens of thousands, a common feature – regardless of identification criteria – is that they cluster around key developmental genes. The broad and overlapping criteria used to define these elements makes it problematic to use the designations applied by the authors of each paper in this review. For clarity, we therefore refer to all regions identified through sequence alignment as CNEs.

ectopic expression was also seen in some cases (e.g. notochord for *sox21*). In another study using mouse embryos, Pennacchio and colleagues identified 167 regions from the human genome that were either conserved between human and fugu, or ultraconserved between human, mouse and rat (Pennacchio et al., 2006). Of these human CNEs, 45% drove expression in discrete anatomical structures, mostly in the nervous system, in mouse embryos at E11.5 (i.e. human sequences tested in mouse, or HM). For example, CNEs upstream of *Sall1* are able to drive reporter gene expression in tissues where *Sall1* expression is normally seen, including in the brain and limb.

These and other studies (e.g. Poulin et al., 2005; Shin et al., 2005) show that a CNE from one species may drive expression in another species, but this does not test whether a CNE drives conserved expression. To test this, and distinguish between any *cis* and/or *trans* changes, the same genomic region should be tested in multiple species or the same CNE from multiple species should be tested in one species (e.g. McEwen et al., 2009; Navratilova et al., 2009; Ritter et al., 2010; Sato et al., 2012). Such studies have shown that although some CNEs drive conserved expression, not all do so. For example, Ritter and colleagues compared the sequence of 875 human CNEs that had been previously tested for expression in mouse (HM) with 151 zebrafish CNEs that had been tested for expression in zebrafish (ZZ) and found that 41 of the sequences aligned. However, when the expression patterns of the HM CNEs were compared with the expression of the ZZ CNEs, only around one-third were found to have similar patterns (Ritter et al., 2010). When these human CNEs were then assayed for their expression in zebrafish (HZ), only four showed the same expression pattern as the cognate zebrafish CNE (ZZ), with the other nine showing a different pattern (Ritter et al., 2010). In other words, although CRMs can be identified through sequence conservation analysis, the expression patterns they drive across species may show little conservation (Fig. 1).

Most studies have concentrated on identifying enhancer activity of CNEs, presumably because this assay leads to an easily observable readout. However, it is possible that some CNEs with no detectable enhancer activity act instead to repress gene expression. This possibility was addressed by Royo and colleagues, who investigated CNEs from zebrafish chromosome 16 in reporter assays (Royo et al., 2011). Thirteen CNEs that did not show enhancer activity in zebrafish transient transgenesis were then tested for their enhancer-blocking ability and three were found to display significant repressor activity (Royo et al., 2011).

Conserved sequences may be sufficient to regulate gene expression, but are they necessary? Although this question has not been widely addressed, one study knocked out four ultraconserved regions that drive gene expression in mouse embryos. The resulting mice showed no observable phenotype, and expression of the genes linked to the conserved elements did not change significantly (Ahituv et al., 2007), suggesting that these conserved regions are not required for the normal activity of the linked genes. However, *Drosophila* studies suggest that phenotypes associated with deletion of a CNE may only appear under conditions of stress because additional 'shadow' enhancers compensate at other times (Frankel et al., 2010; Perry et al., 2010). A more recent study analysed three characterized enhancers that are conserved between human and mouse, and drive gene expression in mouse liver (Patwardhan et al., 2012). Using random mutagenesis and high-throughput sequencing to identify nucleotides that are necessary for robust expression, the authors found that many, but not all, evolutionarily conserved nucleotide mutations affected expression, demonstrating that conserved residues are not necessarily functionally conserved with respect to gene expression (Patwardhan et al., 2012).

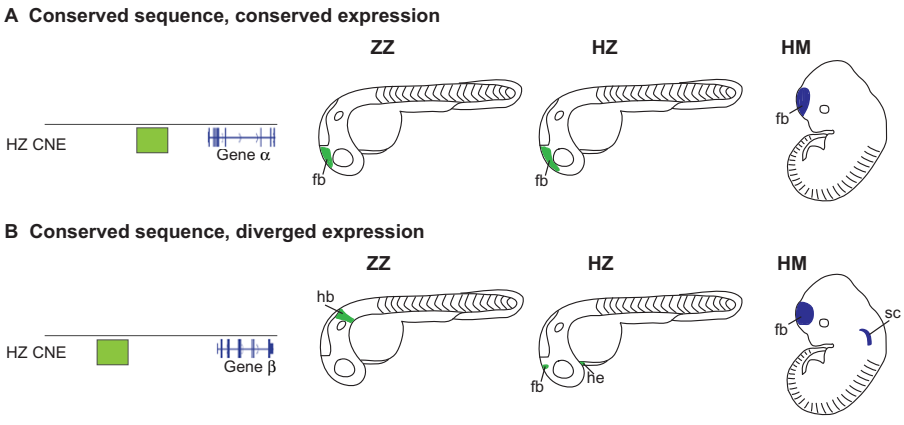

Fig. 1. Conserved regulatory sequences may drive conserved or divergent expression in different species. Conserved non-coding elements (CNEs) between human and zebrafish (HZ CNE) can be used to test enhancer activity. (**A**) An example of a zebrafish CNE tested in zebrafish (ZZ) and the corresponding human CNE tested in zebrafish (HZ) and mouse (HM), all of which show expression in the forebrain (fb). (**B**) Another zebrafish CNE tested in zebrafish (ZZ) and the corresponding human CNE tested in zebrafish (HZ) and mouse (HM), all of which show divergent expression in either the hindbrain (hb), heart (he), forebrain (fb) and/or spinal cord (sc). Images prepared using data from Ritter et al. (Ritter et al., 2010).

There are limitations to the above-described functional assays for testing CNE activity. They may not capture all regulatory activity owing to position effects on the inserted DNA, because heterologous promoters are used in the reporter construct or because only a fraction of developmental time points are assayed, for example. Similarly, the computational method of detecting conserved sequence may miss some regions of conservation. Still, even taking these constraints into account, it is clear that the link between sequence conservation and conservation of gene regulation is inconsistent. Some CNEs can act as CRMs and regulate conserved gene expression whereas other CNEs drive different expression patterns despite highly similar sequence (e.g. Ritter et al., 2010); others are sufficient for gene expression regulation, but not necessary (e.g. Ahituv et al., 2007).

Indeed, there are many important developmental genes that show little evidence for CNEs in the surrounding regulatory DNA (Fig. 2), but that have conserved expression, regulation and function during vertebrate development (e.g. Cutty et al., 2012). As discussed further below, it has also been shown that functional regulatory elements are very flexible and the same gene expression outcome can be brought about by a non-conserved sequence. Therefore, using sequence conservation as the only method to identify regulatory sequence risks overlooking a large part of the regulatory genome. Furthermore, although in many instances CRM sequence and gene expression may remain conserved, some changes in CRM sequence and gene regulation within and between organisms are to be expected. Such changes can account for phenotypic differences between individuals and species, and may be a substrate for evolution leading to speciation (reviewed by Wray, 2007; Wittkopp and Kalay, 2012).

Expression regulation by non sequence-conserved CRMs

An increasing number of studies have identified enhancers (see Glossary, Box 1) that show no overt sequence conservation using usual alignment techniques, but that are able to drive reporter expression in a pattern associated with proximal genes (e.g. Fisher et al., 2006; McGaughey et al., 2009; Blow et al., 2010; Friedli et al., 2010; Chatterjee et al., 2011).

One series of studies, which addresses in detail conserved regulation by non-conserved sequence, has investigated the well-characterized *D. melanogaster even-skipped* (*eve*) control regions (Ludwig and Kreitman, 1995; Ludwig et al., 1998; Ludwig et al., 2000; Hare et al., 2008b). These studies, which compare the control of *eve* across multiple *Drosophila* species and another dipteran family, the Sepsidae, show that the *eve* stripe enhancers from different dipteran species drive conserved expression in *D. melanogaster*, despite overt sequence conservation being low over an extended sequence block. However, short conserved blocks of 20-30 bp are found in these enhancers, and these contain overlapping or adjacent TFBSs. Indeed, the different *eve* enhancers can be identified in different *Drosophila* species by searching for clusters of binding sites for the TFs known to regulate *eve*

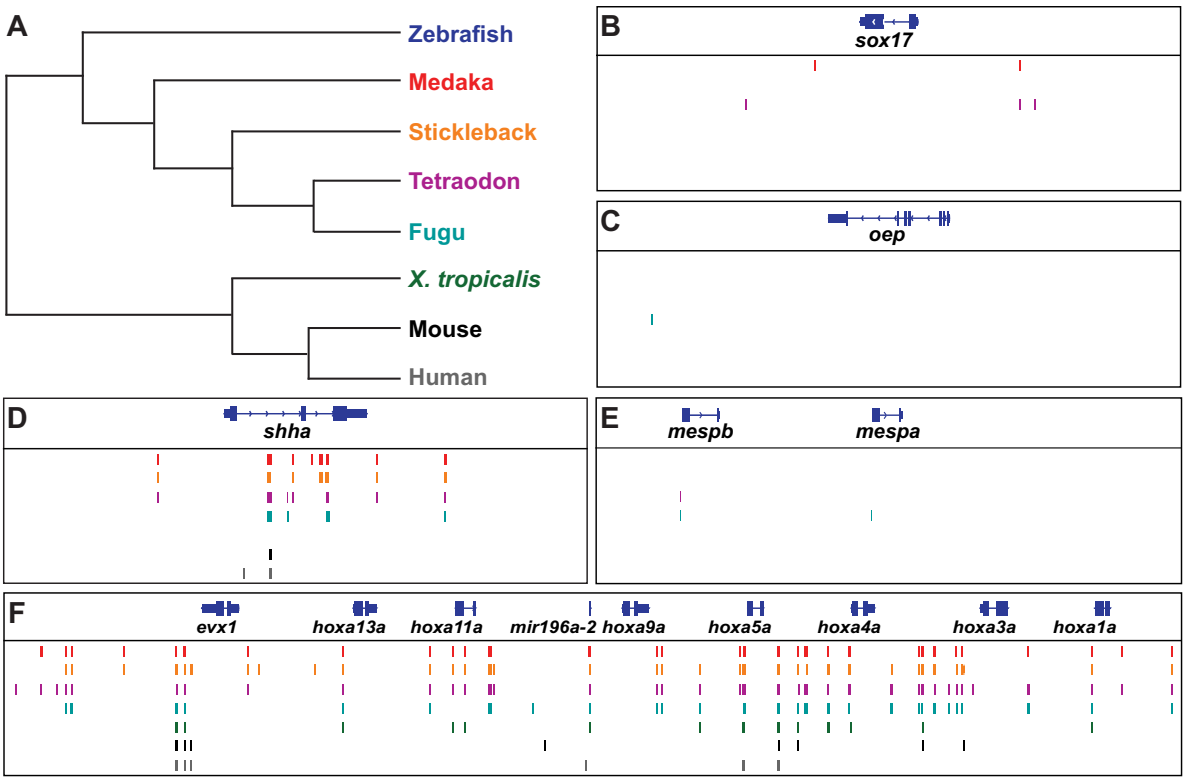

Fig. 2. Examples of conserved non-coding elements surrounding developmental genes. (**A**) Phylogenetic tree indicating the evolutionary relationship between the eight vertebrate species used in this figure. (**B-F**) Genomic windows (30 kb) centred on *sox17* (B), *oep* (C), *shha* (D) and *mespa/b* (E), and a 90 kb genomic window encompassing the *hoxa* gene cluster (F) in zebrafish. Conserved non-coding elements (CNEs) for seven vertebrates species are represented as bars (colour coded as in A) and indicate regions of at least 70% sequence identity over 50 bases between zebrafish and that species (as defined by Engström et al., 2008). Many genes that are essential in vertebrates and have similar regulation [e.g. *sox17*, *oep* (*cripto*), *mespa* and *mespb*] are poorly associated with CNEs. Other essential genes (e.g. *shha*) are highly associated with CNEs in all vertebrate species. Hox gene clusters also contain a large number of CNEs conserved across long evolutionary distances. The number of CNEs shared by species within the same class (medaka, stickleback, fugu, Tetraodon) is greater than within the wider phylum.

expression (Hunchback, Kruppel, Bicoid, Giant, Knirps and Caudal). Overall sequence conservation is low in this case because the distribution and number of sites has changed in the different species, and base substitutions in the TFBSs have also occurred (Hare et al., 2008a; Hare et al., 2008b; but see also Crocker and Erives, 2008). Other comparative studies show a similar outcome – usual alignment techniques do not show conservation but closer inspection of a functionally conserved enhancer reveals known TFBSs, albeit with considerable flexibility in the sequence, spacing, orientation, organization and numbers of these sites (e.g. Romano and Wray, 2003; Oda-Ishii et al., 2005; Cameron and Davidson, 2009).

Although seemingly less common (or perhaps less studied), it is also possible for the regulatory sequence for a specific gene present in two related species to contain entirely different TFBSs but still result in the same pattern of gene expression when tested in those organisms (e.g. Takahashi et al., 1999; Dayal et al., 2004). For example, the *brachyury* gene is expressed in the notochord of ascidian embryos. The CRMs required for this expression have been characterized in both *Ciona intestinalis* and *Halocynthia roretzi*, and each CRM can drive reporter gene expression in the notochord of both species (Takahashi et al., 1999). However, not only is there no detectable sequence conservation, but the TFBSs present are also different: the *C. intestinalis* CRM has sites for SuH in the activating region and an additional region that mediates negative regulation of expression, whereas the *H. roretzi* CRM has a T-box binding site in the activating region and no repressive sites (Takahashi et al., 1999). These studies show, then, that a sequence with little or no conservation, even a sequence with different TFBSs, is able to drive conserved expression across species in some situations.

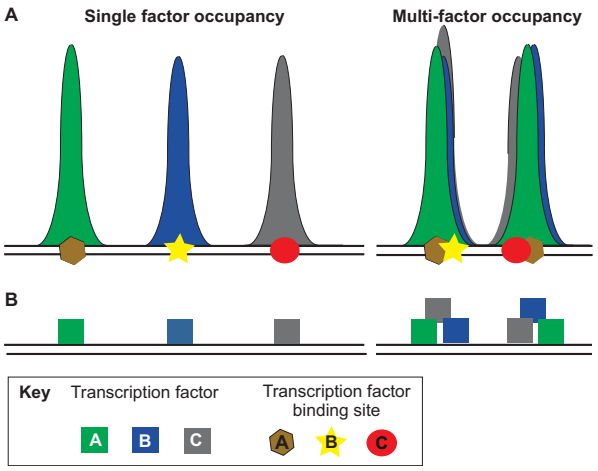

Fig. 3. Transcription factor-binding site incidence in *cis*-regulatory modules may not be completely predictive of operative transcription factors. (**A**) ChIP-seq peaks of binding for transcription factors (TFs) A (green), B (blue) and C (grey) over identified transcription factor-binding sites (TFBSs) for each factor (brown hexagon, yellow star and red circle) on DNA (double black line). Single TFs that bind DNA directly in the presence of their TFBS (single factor occupancy) may also bind DNA via other TFs, and in the absence of their own DNA recognition sequence (multi-factor occupancy). (**B**) The arrangement of individual TFs on the DNA for the example in A. For multi-factor occupancy, a TFBS may be dispensable if the cognate TF interacts with the *cis*-regulatory module (CRM) via other TFs. Knowledge of TFs that regulate a gene therefore does not necessarily facilitate sequence-based CRM identification. Equally, where a CRM has been directly identified, determining the operative TFs may not be comprehensively achieved based on the constituent TFBSs.

Alternative methods of CRM identification

Covert conservation and lessons from multi-factorial studies

The majority of studies discussed thus far have used sequence-based approaches involving multi-species sequence alignment to identify CRMs, but, as already noted, some functionally conserved CRMs may contain clusters of TFBSs without overt sequence conservation. Thus, it may be more appropriate to search for this 'covert' conservation, i.e. the conserved proximal incidence of TFBSs regardless of the overall conservation of the region (e.g. Taher et al., 2011). Gene regulation through individual CRMs tends to involve binding of multiple TFs, either cooperatively or independently. Assuming this binding is manifested at the sequence level, such as by the presence and consistent spacing of known TFBSs, it should be possible to identify CRMs.

Clear examples of grammatical rules (see Glossary, Box 1) that define certain enhancers do exist and could be used to identify CRMs in some instances (Senger et al., 2004; Panne, 2008). However, such an approach requires prior knowledge of the factors that regulate a gene of interest, the sequence motifs to which those TFs bind and an appropriate genomic search space. The use of sequence to identify CRMs is further complicated because not all CRMs are defined by a detectable architecture. This complexity is illustrated in a recent study in *Drosophila* (Junion et al., 2012). In this work, the authors performed chromatin immunoprecipitation (ChIP; discussed in more detail below) combined with genomic microarrays for five TFs expressed in the cardiac mesoderm, followed by sequence analysis of regions bound by these factors both individually and in combination. They found that, whereas genomic regions occupied by the individual factors were significantly enriched for the TFBS of that factor, sites occupied by multiple factors lacked TFBSs for some of those factors (Fig. 3). Thus, a TF capable of binding DNA can be recruited to genomic locations lacking its TFBS via physical interactions with other DNA-bound TFs. Furthermore, no coherent grammar could be detected for the motifs that were present (Junion et al., 2012). These are interesting results as they suggest that many functional CRMs may be resistant to indirect identification, owing to their liberal configuration, and that information about factors acting through CRMs may also be absent, owing to their indirect genomic contact. Indeed, the absence of detectable grammar may be the most common situation since regulatory elements emerged during the stochastic process of evolution (Wray et al., 2003), resulting in CRMs without an easily recognizable pattern but with consistencies such as the binding of key TFBSs.

To summarize, without a clear idea of what a CRM looks like in terms of grammar, architecture and sequence conservation, *in silico* prediction is currently not straightforward. A superior alternative to identify CRMs, then, could be the use of direct methods, such as DNase hypersensitivity and ChIP for histone modifications and specific TFs. This, in turn, should give us improved information on how different CRMs are assembled and therefore a better understanding of CRM grammar, where it exists.

DNase hypersensitivity and ChIP-seq – direct approaches to CRM identification

The innovation of next generation sequencing has inevitably led to the increased practicality and incidence of genome-scale approaches to study *cis* regulation and TF function. DNaseI

hypersensitive sites are associated with enhancers, promoters, silencers and insulators (see Glossary, Box 1) and therefore DNase hypersensitivity combined with deep sequencing has been used to map active regulatory regions (see Glossary, Box 1) across the genome in multiple cell lines (Gross and Garrard, 1988; Gaszner and Felsenfeld, 2006; Boyle et al., 2008; Song et al., 2011; Neph et al., 2012; Thurman et al., 2012). Although whole-genome DNase hypersensitivity assays have yet to be applied to developing embryos, it is probable that some cell line information can be extrapolated to the regulation of embryonic genes. In a recent study, McBride and colleagues identified 50 DNase hypersensitive sites in mouse cell lines across the control region of Pax6 (McBride et al., 2011). The majority of these are yet to be tested for function, but 11 are known enhancers and another four that were tested in this study showed enhancer activity in the embryo. Moreover, the pattern of cleavage at DNase hypersensitive sites can be used to map TF occupancy at nucleotide resolution, and *de novo* motif searches enable the identification of TFs that may bind and regulate that region (Neph et al., 2012).

Alternatively, genome-wide ChIP studies have related methylation and acetylation of histones to sites of active transcription and repression (Pokholok et al., 2005; Bernstein et al., 2006), 'poised' promoters and enhancers (see Glossary, Box 1) (Creyghton et al., 2010; Rada-Iglesias et al., 2011) and also to distal enhancers (Heintzman et al., 2007), whereas general TFs can be used to identify active promoters and enhancers (Kim et al., 2005; Heintzman et al., 2007; Visel et al., 2009). ChIP-seq for specific TFs allows *de novo* discovery of TF-binding motifs, both for the ChIPed factor and potential interacting factors. However, it is clear that TF binding alone is not always a functional event and does not necessarily directly reveal a transcriptional target gene (discussed by Farnham, 2009). Instead, colocalization with other markers, such as histone modifications, RNA polymerase II occupancy or other cooperating TFs, has proved a successful means of identifying functional TF-binding events (e.g. Kim et al., 2005; Rada-Iglesias et al., 2011; Junion et al., 2012). Although this combined approach might be straightforward when starting material is not limited, this is rarely the case when performing *in vivo* experiments in a developmental context. Functional binding of a specific TF may be inferred by other means, however, such as demonstrating that the expression level of a gene with proximal TF binding is dependent upon that factor – by integrating ChIP-seq data with RNA-seq or microarray data for loss or gain of function of the immunoprecipitated TF. This method highlights which genes with proximal TF binding require that binding for their correct regulation. A particularly relevant example of this is the approach taken by Kunarso et al., where microarray analysis of RNAi knockdown of key pluripotency TFs in human and mouse embryonic stem (ES) cells was integrated with ChIP-seq data for the same TFs (Kunarso et al., 2010). This approach revealed that ~25% of putative functional CRMs originate from species-specific transposable elements, which lead to substantial remodelling of the transcriptional circuitry of ES cells between mouse and human.

Developmental and evolutionary conservation of CRMs

Although there are plenty of studies that identify active regulatory sequences using ChIP approaches, they tend to be limited to a single biological context in a single species. Studies that comprehensively address CRM use in multiple species or in multiple contexts within one species are rare; to our knowledge, there are no published studies that do both. For these reasons, we lack a clear picture of the extent to which CRM conservation is relevant across developmental time and space, as well as across evolution. In addition, few TF ChIP-seq studies analyse the sequence of bound regions beyond identifying the binding element of the factor; the studies that do, provide the most information about conservation and *cis* regulation, and we discuss some of these in more detail below.

Lessons from pan-evolutionary studies

One key example of a study that tested the degree to which conserved sequences are consistently used throughout evolution involved ChIP-seq for the mesodermal TF Twist in gastrulas of six *Drosophila* species (He et al., 2011). They applied an approach using liftover – a tool that identifies equivalent genomic regions between species using conserved sequences as landmarks (Kent et al., 2002). In so doing, they could test whether TF binding consistently occurs in corresponding regions between species, regardless of whether the sequence of these regions is conserved. The authors showed that Twist generally binds the equivalent genomic locations in different species. Crucially, they also showed that sequence in these regions is not necessarily highly conserved. However, conserved binding events are highly enriched for high quality TFBSs compared with species-specific binding events, both for Twist and its known partner TFs. The mechanism for conserved binding therefore appears to be conservation of TFBSs in proximity to genes, rather than conservation of larger-scale stretches of sequences (Fig. 4). Furthermore, the conserved binding events are significantly associated with genes that change in response to Twist loss of function. The implication of this finding is clear – ChIP-seq approaches can identify CRMs that are conserved in a manner that is functionally significant, whereas searching for CRMs by conventional genomic sequence comparison approaches would not have identified these regions.

Another study applying ChIP-seq for two different TFs (CCAAT/enhancer-binding protein α and hepatocyte nuclear factor 4 α) in the livers of five vertebrates species provides an alternative view of conservation of TF-binding events (Schmidt et al., 2010). This work demonstrated that the majority of binding events are species specific, rather than consistently localized in conserved regions. Binding to conserved sequences in one species was rarely indicative of binding to the homologous sequence in others. These differences in binding were consistently observed between human and mouse in the livers of both species, and also in the livers of aneuploid mice harbouring human chromosome 21. Binding to the human chromosome in mouse was representative of binding to the endogenous chromosome in human, rather than binding to mouse chromosomes (Wilson et al., 2008) (Fig. 4). The differences in binding between species are therefore unlikely to be due to non-equivalence in the assayed tissue. Similarly, Kunarso and colleagues found that whereas genes with proximal binding of OCT4 and NANOG in human and mouse ES cells were substantially similar, only ~5% of sites were homologously occupied in both species (Kunarso et al., 2010). These studies are compelling, as they suggest that enhancer prediction or direct identification in a single species does not necessarily reveal a general mechanism of regulation for the associated gene.

In another recent study, Schmidt et al. identified the genomic binding regions of two TFs (CCAAT/enhancer binding protein α and peroxisome proliferator activated receptor γ) in human and mouse adipocytes using ChIP-seq (Schmidt et al., 2011). Their results complement the previously mentioned studies; they too

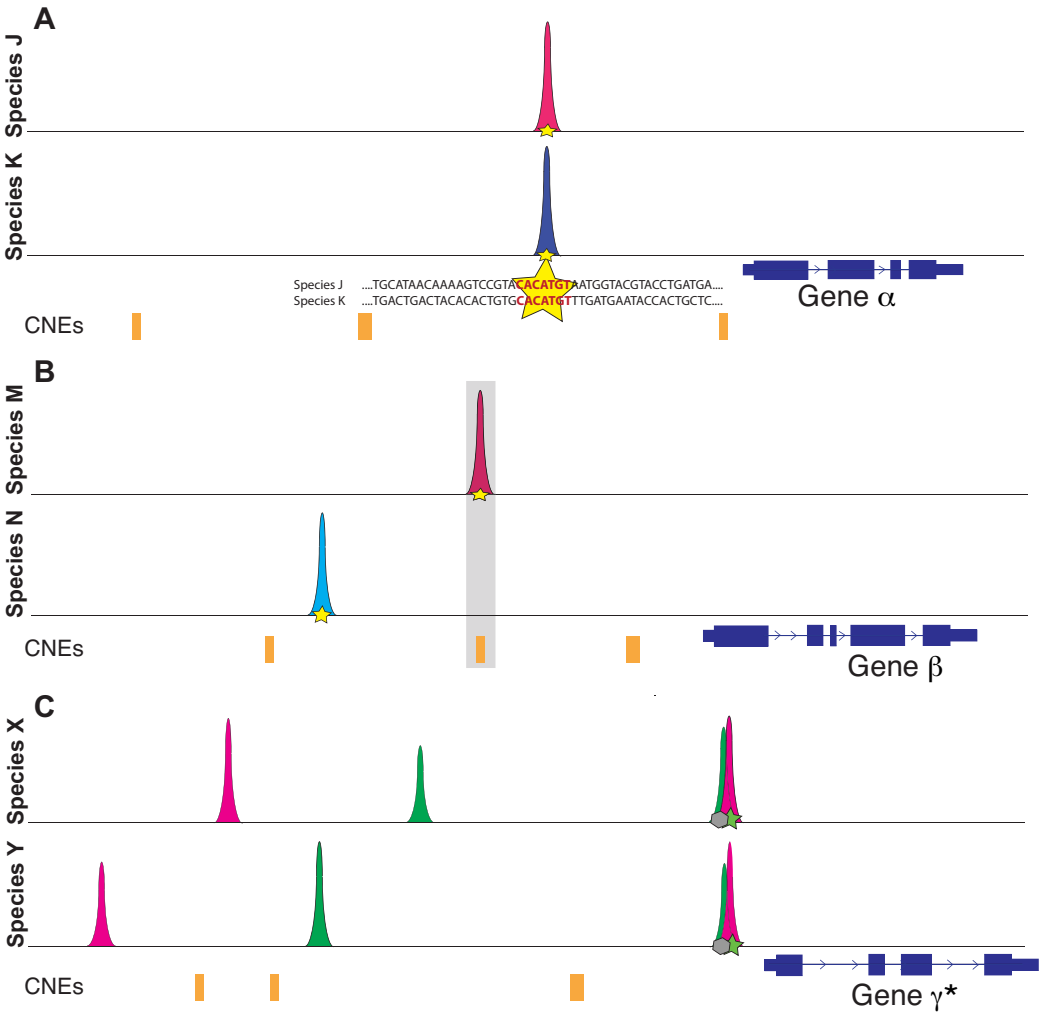

Fig. 4. Transcription factor binding across evolution is strongly correlated with incidence of specific transcription factor-binding sites, but not broader regions of conservation. Multiple studies have addressed the extent to which transcription factor (TF)-binding events are conserved across evolutionary time and the sequence characteristics associated with conserved binding. Three notable findings are depicted: (**A**) the genomic locations of TF binding (coloured peaks) relative to genes are conserved, but *cis*-regulatory module (CRM) sequence is not highly conserved beyond the presence of relatively short transcription factor-binding sites (TFBSs) (yellow stars) – these CRMs are not conserved non-coding elements (CNEs); (**B**) TF binding to CNEs in one species is not predictive of binding to the orthologous CNE in a related species (grey box); and (**C**) co-binding of TFs at high-quality TFBSs (grey hexagon, green star) in close proximity to genes is most associated with regulated genes (*). Distal binding tends to consist of individual factors and is less likely to be functional. CNEs are shown as orange boxes.

found that most binding events were species specific, but that liftover revealed a minority of common binding regions. These regions were characterized by co-occupancy of the two factors (and their TFBSs) and proximity to genes upregulated during adipogenesis. Conserved binding events therefore bear the hallmarks of functionality.

There are substantial differences in population genetics between *Drosophila* and mammals that are likely to influence the organization and complexity of CRMs. Specifically, the difference in gene number and genome size, and the intensity of purifying selection owing to population size may restrict the variation in *cis*-regulatory mechanisms in *Drosophila* compared with mammals (Adams et al., 2000; Lander et al., 2001; Lynch and Conery, 2003). In addition, the aforementioned ChIP studies focus on different classes of TF that may be subject to different restrictions in CRM evolution and use. *Drosophila* Twist, for example, is a master regulator of embryonic morphogenesis and may therefore exhibit less divergent binding than less critical morphogenic factors, such as those involved in mammalian adipogenesis.

The CRMs identified in these studies, although often conserved at the level of TFBSs, are not CNEs. If highly conserved sequences are capable of driving expression in multiple species but are not used to doing so, it is not clear why they are conserved. There are three possibilities: (1) that they perform a function other than the identified *cis* regulation; (2) that the functional analyses of these sequences are not appropriate for clarifying their *cis*-regulatory function; or (3) that redundancy of CRMs leads to their conservation without necessitating consistent use (Hong et al., 2008).

A common feature of the aforementioned studies is their restriction to a static biological context, but there are many examples in developmental biology where TF binding and/or functional use of enhancers changes over time (Garber et al., 2012; Jakobsen et al., 2007; Sandmann et al., 2006; Sudou et al., 2012; Wilczyński and Furlong, 2010). In view of this, it is possible that

CNEs are functional CRMs but are not universally used during development. A broader view would therefore be required to capture their function and trends in their use.

Lessons from pan-developmental studies

The use of conserved non-coding sequences throughout the dynamic process of development is only now being explored. In a recent study by Bogdanovic et al. (Bogdanovic et al., 2012), the changing landscape of histone modification, which indicates enhancer use, was assayed at four stages during zebrafish embryogenesis. A distinct caveat of genomics approaches applied to whole embryos is that a heterogeneous cell population is analysed *en masse*. Consequently, it is difficult to deconvolute the resulting signal to reveal the precise spatial use of enhancers. Nevertheless, the authors noticed a significant increase in the degree of sequence conservation of active enhancers at the end of gastrulation. This corresponds to the onset of the controversial 'phylotypic stage' – a stage at which the phenotypic diversity within an evolutionary group is thought to be reduced (see Glossary, Box 1). Although it has not been determined whether this signal represents a property of the embryos as a whole, or specific cell populations, it may suggest that conservation of CRMs occurs to regulate expression at such a crucial stage – where stringent regulation of developmental genes is required to ensure a viable body plan. Precise regulation of gene expression under other conditions, such as in the liver of adult vertebrates, may require less precise regulation and therefore less conserved CRMs.

Although such a theory is unproven, there are two recent gene expression studies that support it. Developmental time courses of expression were performed in six *Drosophila* (Kalinka et al., 2010) and five *Caenorhabditis* species (Levin et al., 2012), spanning around 30-40 million years of evolution – the equivalent range represented in the vertebrate liver study. Both studies identified stages of significantly less divergent gene expression corresponding to the phylotypic stage in each evolutionary group. They also observed that the most consistent expression at these stages was among key developmental genes, rather than among genes controlling more specialized non-developmental processes, such as the immune response. As previously mentioned, CNEs are more closely associated with key developmental genes. It is therefore feasible that use of highly conserved *cis*-regulatory sequence coincides with the tighter regulation of the expression of these genes at the phylotypic stage. Other studies have suggested that the transcriptome at the phylotypic stage is also the most ancient in terms of the evolutionary emergence of the genes expressed, both in animals and plants (Domazet-Lošo and Tautz, 2010; Quint et al., 2012).

Conclusions and future directions

CRM identification

The numerous different approaches that have been used to identify CRMs have met with variable success but have taught us much about the makeup of CRMs, as well as posing further questions. There appear to be two distinct classes of CRM at the sequence level: those that are highly conserved and are detectable by multi-species genomic alignments; and those defined only by the presence of numerous TFBSs. The second class is particularly difficult to identify as our understanding of the make-up of such CRMs is limited. Some have a rigid and discernible grammar, characterized by consistent location and orientation of TFBSs, which aids in their identification (Papatsenko et al., 2009). This does not appear to be universal, however, and prior knowledge of TFBS sequences is often required to identify a functional module. Other CRMs may not contain TFBSs for all of the expected binding factors, potentially hindering their identification even where the *cis*-acting factors are known (Junion et al., 2012).

Direct methods of identifying CRMs using ChIP and DNase hypersensitivity approaches allow the identification of CRMs that could not be found by sequence analysis alone (e.g. Blow et al., 2010). However, such direct methods also have their drawbacks as they require availability of appropriate antibodies and access to sufficient amounts of appropriate material, conditions that are not always available or practical for developmental model organisms.

Whether identified directly or indirectly, appropriate characterization of a putative CRM is crucial. Assumptions regarding the properties of a CRM cannot be safely made. This is because sequence conservation is not always associated with functional conservation (e.g. Ritter et al., 2010). Furthermore, TF binding is not always a functional transcriptional event, nor is it straightforward to correlate binding with the relevant target gene even when the binding is functional (Farnham, 2009). Although our ability to identify CRMs is improving all the time, the rigorous study of the regulation of any given gene is still not a trivial undertaking.

A possible function of highly conserved sequence

As highlighted above, it appears that any correlation between CNEs and *cis* regulation is context and stage specific. The limited available studies suggest that sequence conservation at active enhancers is stage specific (Bogdanovic et al., 2012), as is the high correlation of orthologous gene expression (Kalinka et al., 2010; Levin et al., 2012); both events seem to converge during the phylotypic stage. One credible hypothesis may be that highly conserved non-coding sequence is responsible for ensuring the similarity of gene expression and body pattern within phyla at this time point. Such stringent gene regulation may not be required outside of the phylotypic stage and so highly conserved regions are less likely to be functional at other times, and less stringently conserved mechanisms are used to control gene expression. Such a model was proposed by Denis Duboule as early as 1994 (Duboule, 1994), noting the striking requirement for stringent sequential activation of vertebrate Hox genes during the phylotypic progression. Interestingly vertebrate Hox gene clusters contain large numbers of CNEs (Fig. 2). It is possible that CNEs reflect the need for tight regulation of a particular set of genes during a specific developmental period.

However, such a hypothesis has yet to be properly tested, and represents a considerable undertaking. In order to determine whether CNEs are associated with phylotypic gene regulation, it would be necessary to perform ChIP-seq for modified histones and general transcription factors to identify active enhancers and promoters in multiple species at multiple equivalent developmental stages. This would be carried out to compare non-phylotypic and phylotypic time points. To demonstrate a link between chromatin marks at CNEs and conserved gene expression, it would also be necessary to perform concomitant expression analyses at these stages. The hypothesis is that functional CRMs identified by ChIP-seq would show greater association with CNEs during phylotypic stages in each species, and that gene expression would also be less diverse at these time points. If this were so, it would represent a significant step forward in our understanding of transcriptional regulation and the role of conserved non-coding sequences in developmental biology.

Looking forward

As discussed in this Review, although recent studies have provided tantalizing clues about the relationship between conservation and

cis regulation during development, we are still a long way from seeing the full picture. However, the growing number of sequenced genomes should improve our ability to identify less stringently conserved genomic regions that may be CRMs. The increased cataloguing of functionally active genomic regions through DNase I hypersensitivity and ChIP assays coupled with sequence analysis is also likely to reveal *cis*-regulatory grammar in more detail and will, in turn, increase the accuracy of indirect approaches of CRM identification. With these high-throughput and computational methods, and further functional expression analyses, there is realistic hope for rapid progress in our understanding of *cis* regulation during embryonic development and the extent to which conserved sequences drive gene expression.

Acknowledgements
We thank members of the Wardle lab, Matt Clark, Cathy Wilson and unknown reviewers for their helpful comments on the manuscript.

Funding
F.C.W. is supported by a Medical Research Council Career Development Award and Lister Institute Research Prize.

Competing interests statement
The authors declare no competing financial interests.

References

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F. et al. (2000). The genome sequence of Drosophila melanogaster. *Science* **287**, 2185-2195.

Ahituv, N., Zhu, Y., Visel, A., Holt, A., Afzal, V., Pennacchio, L. A. and Rubin, E. M. (2007). Deletion of ultraconserved elements yields viable mice. *PLoS Biol.* **5**, e234.

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J. M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A. et al. (2002). Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. *Science* **297**, 1301-1310.

Bejerano, G., Pheasant, M., Makunin, I., Stephen, S., Kent, W. J., Mattick, J. S. and Haussler, D. (2004). Ultraconserved elements in the human genome. *Science* **304**, 1321-1325.

Bernstein, B. E., Mikkelsen, T. S., Xie, X., Kamal, M., Huebert, D. J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K. et al. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315-326.

Blow, M. J., McCulley, D. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F. et al. (2010). ChIP-Seq identification of weakly conserved heart enhancers. *Nat. Genet.* **42**, 806-810.

Bogdanovic, O., Fernandez-Miñán, A., Tena, J. J., de la Calle-Mustienes, E., Hidalgo, C., van Kruysbergen, I., van Heeringen, S. J., Veenstra, G. J. and Gómez-Skarmeta, J. L. (2012). Dynamics of enhancer chromatin signatures mark the transition from pluripotency to cell specification during embryogenesis. *Genome Res.* **22**, 2043-2053.

Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., Furey, T. S. and Crawford, G. E. (2008). High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**, 311-322.

C. elegans Sequencing Consortium (1998). Genome sequence of the nematode C. elegans: a platform for investigating biology. *Science* **282**, 2012-2018.

Cameron, R. A. and Davidson, E. H. (2009). Flexibility of transcription factor target site position in conserved cis-regulatory modules. *Dev. Biol.* **336**, 122-135.

Chatterjee, S., Bourque, G. and Lufkin, T. (2011). Conserved and non-conserved enhancers direct tissue specific transcription in ancient germ layer specific developmental control genes. *BMC Dev. Biol.* **11**, 63.

Clarke, S. L., VanderMeer, J. E., Wenger, A. M., Schaar, B. T., Ahituv, N. and Bejerano, G. (2012). Human developmental enhancers conserved between deuterostomes and protostomes. *PLoS Genet.* **8**, e1002852.

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., Hanna, J., Lodato, M. A., Frampton, G. M., Sharp, P. A. et al. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. USA* **107**, 21931-21936.

Crocker, J. and Erives, A. (2008). A closer look at the eve stripe 2 enhancers of Drosophila and Themira. *PLoS Genet.* **4**, e1000276.

Cutty, S. J., Fior, R., Henriques, P. M., Saúde, L. and Wardle, F. C. (2012). Identification and expression analysis of two novel members of the Mesp family in zebrafish. *Int. J. Dev. Biol.* **56**, 285-294.

Davidson, E. H. (2006). *The Regulatory Genome: Gene Regulatory Networks in Development and Evolution*. Amsterdam, The Netherlands: Academic Press.

Dayal, S., Kiyama, T., Villinski, J. T., Zhang, N., Liang, S. and Klein, W. H. (2004). Creation of cis-regulatory elements during sea urchin evolution by co-option and optimization of a repetitive sequence adjacent to the spec2a gene. *Dev. Biol.* **273**, 436-453.

Dermitzakis, E. T. and Clark, A. G. (2002). Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.* **19**, 1114-1121.

Domazet-Lošo, T. and Tautz, D. (2010). A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. *Nature* **468**, 815-818.

Duboule, D. (1994). Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate Bauplan and the evolution of morphologies through heterochrony. *Development Suppl.* **1994**, 135-142.

Elgar, G. (2009). Pan-vertebrate conserved non-coding sequences associated with developmental regulation. *Brief. Funct. Genomic Proteomics* **8**, 256-265.

Engström, P. G., Ho Sui, S. J., Drivenes, O., Becker, T. S. and Lenhard, B. (2007). Genomic regulatory blocks underlie extensive microsynteny conservation in insects. *Genome Res.* **17**, 1898-1908.

Engström, P. G., Fredman, D. and Lenhard, B. (2008). Ancora: a web resource for exploring highly conserved noncoding elements and their association with developmental regulatory genes. *Genome Biol.* **9**, R34.

Farnham, P. J. (2009). Insights from genomic profiling of transcription factors. *Nat. Rev. Genet.* **10**, 605-616.

Fisher, S., Grice, E. A., Vinton, R. M., Bessling, S. L. and McCallion, A. S. (2006). Conservation of RET regulatory function from human to zebrafish without sequence similarity. *Science* **312**, 276-279.

Frankel, N., Davis, G. K., Vargas, D., Wang, S., Payre, F. and Stern, D. L. (2010). Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature* **466**, 490-493.

Friedli, M., Barde, I., Arcangeli, M., Verp, S., Quazzola, A., Zakany, J., Lin-Marq, N., Robyr, D., Attanasio, C., Spitz, F. et al. (2010). A systematic enhancer screen using lentivector transgenesis identifies conserved and non-conserved functional elements at the Olig1 and Olig2 locus. *PLoS ONE* **5**, e15741.

Garber, M., Yosef, N., Goren, A., Raychowdhury, R., Thielke, A., Guttman, M., Robinson, J., Minie, B., Chevrier, N., Itzhaki, Z. et al. (2012). A high-throughput chromatin immunoprecipitation approach reveals principles of dynamic gene regulation in mammals. *Mol. Cell* **47**, 810-822.

Gaszner, M. and Felsenfeld, G. (2006). Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat. Rev. Genet.* **7**, 703-713.

Glazov, E. A., Pheasant, M., McGraw, E. A., Bejerano, G. and Mattick, J. S. (2005). Ultraconserved elements in insect genomes: a highly conserved intronic sequence implicated in the control of homothorax mRNA splicing. *Genome Res.* **15**, 800-808.

Gordon, K. L. and Ruvinsky, I. (2012). Tempo and mode in evolution of transcriptional regulation. *PLoS Genet.* **8**, e1002432.

Gross, D. S. and Garrard, W. T. (1988). Nuclease hypersensitive sites in chromatin. *Annu. Rev. Biochem.* **57**, 159-197.

Hare, E. E., Peterson, B. K. and Eisen, M. B. (2008a). A careful look at binding site reorganization in the even-skipped enhancers of Drosophila and sepsids. *PLoS Genet.* **4**, e1000268.

Hare, E. E., Peterson, B. K., Iyer, V. N., Meier, R. and Eisen, M. B. (2008b). Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. *PLoS Genet.* **4**, e1000106.

He, Q., Bardet, A. F., Patton, B., Purvis, J., Johnston, J., Paulson, A., Gogol, M., Stark, A. and Zeitlinger, J. (2011). High conservation of transcription factor binding and evidence for combinatorial regulation across six Drosophila species. *Nat. Genet.* **43**, 414-420.

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., Barrera, L. O., Van Calcar, S., Qu, C., Ching, K. A. et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39**, 311-318.

Hong, J. W., Hendrix, D. A. and Levine, M. S. (2008). Shadow enhancers as a source of evolutionary novelty. *Science* **321**, 1314.

Jakobsen, J. S., Braun, M., Astorga, J., Gustafson, E. H., Sandmann, T., Karzynski, M., Carlsson, P. and Furlong, E. E. (2007). Temporal ChIP-on-chip reveals Biniou as a universal regulator of the visceral muscle transcriptional network. *Genes Dev.* **21**, 2448-2460.

Johnson, D. S., Davidson, B., Brown, C. D., Smith, W. C. and Sidow, A. (2004). Noncoding regulatory sequences of Ciona exhibit strong correspondence between evolutionary constraint and functional importance. *Genome Res.* **14**, 2448-2456.

Junion, G., Spivakov, M., Girardot, C., Braun, M., Gustafson, E. H., Birney, E. and Furlong, E. E. (2012). A transcription factor collective defines cardiac cell fate and reflects lineage history. *Cell* **148**, 473-486.

Kalinka, A. T., Varga, K. M., Gerrard, D. T., Preibisch, S., Corcoran, D. L., Jarrells, J., Ohler, U., Bergman, C. M. and Tomancak, P. (2010). Gene expression divergence recapitulates the developmental hourglass model. *Nature* **468**, 811-814.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M. and Haussler, D. (2002). The human genome browser at UCSC. *Genome Res.* **12**, 996-1006.
Kikuta, H., Laplante, M., Navratilova, P., Komisarczuk, A. Z., Engström, P. G., Fredman, D., Akalin, A., Caccamo, M., Sealy, I., Howe, K. et al. (2007). Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res.* **17**, 545-555.
Kim, T. H., Barrera, L. O., Zheng, M., Qu, C., Singer, M. A., Richmond, T. A., Wu, Y., Green, R. D. and Ren, B. (2005). A high-resolution map of active promoters in the human genome. *Nature* **436**, 876-880.
Kunarso, G., Chia, N. Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y. S., Ng, H. H. and Bourque, G. (2010). Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat. Genet.* **42**, 631-634.
Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W. et al. (2001). Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921.
Levin, M., Hashimshony, T., Wagner, F. and Yanai, I. (2012). Developmental milestones punctuate gene expression in the Caenorhabditis embryo. *Dev. Cell* **22**, 1101-1108.
Li, Q., Ritter, D., Yang, N., Dong, Z., Li, H., Chuang, J. H. and Guo, S. (2010). A systematic approach to identify functional motifs within vertebrate developmental enhancers. *Dev. Biol.* **337**, 484-495.
Ludwig, M. Z. and Kreitman, M. (1995). Evolutionary dynamics of the enhancer region of even-skipped in Drosophila. *Mol. Biol. Evol.* **12**, 1002-1011.
Ludwig, M. Z., Patel, N. H. and Kreitman, M. (1998). Functional analysis of eve stripe 2 enhancer evolution in Drosophila: rules governing conservation and change. *Development* **125**, 949-958.
Ludwig, M. Z., Bergman, C., Patel, N. H. and Kreitman, M. (2000). Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* **403**, 564-567.
Lynch, M. and Conery, J. S. (2003). The origins of genome complexity. *Science* **302**, 1401-1404.
McBride, D. J., Buckle, A., van Heyningen, V. and Kleinjan, D. A. (2011). DNaseI hypersensitivity and ultraconservation reveal novel, interdependent long-range enhancers at the complex Pax6 cis-regulatory region. *PLoS ONE* **6**, e28616.
McEwen, G. K., Goode, D. K., Parker, H. J., Woolfe, A., Callaway, H. and Elgar, G. (2009). Early evolution of conserved regulatory sequences associated with development in vertebrates. *PLoS Genet.* **5**, e1000762.
McGaughey, D. M., Stine, Z. E., Huynh, J. L., Vinton, R. M. and McCallion, A. S. (2009). Asymmetrical distribution of non-conserved regulatory sequences at PHOX2B is reflected at the ENCODE loci and illuminates a possible genome-wide trend. *BMC Genomics* **10**, 8.
McPherson, J. D., Marra, M., Hillier, L., Waterston, R. H., Chinwalla, A., Wallis, J., Sekhon, M., Wylie, K., Mardis, E. R., Wilson, R. K. et al. (2001). A physical map of the human genome. *Nature* **409**, 934-941.
Narayanan, A. and Lekven, A. C. (2012). Biphasic wnt8a expression is achieved through interactions of multiple regulatory inputs. *Dev. Dyn.* **241**, 1062-1075.
Navratilova, P., Fredman, D., Hawkins, T. A., Turner, K., Lenhard, B. and Becker, T. S. (2009). Systematic human/zebrafish comparative identification of cis-regulatory activity around vertebrate developmental transcription factor genes. *Dev. Biol.* **327**, 526-540.
Neph, S., Vierstra, J., Stergachis, A. B., Reynolds, A. P., Haugen, E., Vernot, B., Thurman, R. E., John, S., Sandstrom, R., Johnson, A. K. et al. (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83-90.
Nobrega, M. A., Ovcharenko, I., Afzal, V. and Rubin, E. M. (2003). Scanning human gene deserts for long-range enhancers. *Science* **302**, 413.
Oda-Ishii, I., Bertrand, V., Matsuo, I., Lemaire, P. and Saiga, H. (2005). Making very similar embryos with divergent genomes: conservation of regulatory mechanisms of Otx between the ascidians Halocynthia roretzi and Ciona intestinalis. *Development* **132**, 1663-1674.
Panne, D. (2008). The enhanceosome. *Curr. Opin. Struct. Biol.* **18**, 236-242.
Papatsenko, D., Goltsev, Y. and Levine, M. (2009). Organization of developmental enhancers in the Drosophila embryo. *Nucleic Acids Res.* **37**, 5665-5677.
Patwardhan, R. P., Hiatt, J. B., Witten, D. M., Kim, M. J., Smith, R. P., May, D., Lee, C., Andrie, J. M., Lee, S. I., Cooper, G. M. et al. (2012). Massively parallel functional dissection of mammalian enhancers in vivo. *Nat. Biotechnol.* **30**, 265-270.
Pennacchio, L. A., Ahituv, N., Moses, A. M., Prabhakar, S., Nobrega, M. A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K. D. et al. (2006). In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**, 499-502.
Perry, M. W., Boettiger, A. N., Bothma, J. P. and Levine, M. (2010). Shadow enhancers foster robustness of Drosophila gastrulation. *Curr. Biol.* **20**, 1562-1567.
Pokholok, D. K., Harbison, C. T., Levine, S., Cole, M., Hannett, N. M., Lee, T. I., Bell, G. W., Walker, K., Rolfe, P. A., Herbolsheimer, E. et al. (2005). Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell* **122**, 517-527.
Poulin, F., Nobrega, M. A., Plajzer-Frick, I., Holt, A., Afzal, V., Rubin, E. M. and Pennacchio, L. A. (2005). In vivo characterization of a vertebrate ultraconserved enhancer. *Genomics* **85**, 774-781.
Quint, M., Drost, H. G., Gabel, A., Ullrich, K. K., Bönn, M. and Grosse, I. (2012). A transcriptomic hourglass in plant embryogenesis. *Nature* **490**, 98-101.
Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S. A., Flynn, R. A. and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. *Nature* **470**, 279-283.
Ritter, D. I., Li, Q., Kostka, D., Pollard, K. S., Guo, S. and Chuang, J. H. (2010). The importance of being cis: evolution of orthologous fish and mammalian enhancer activity. *Mol. Biol. Evol.* **27**, 2322-2332.
Romano, L. A. and Wray, G. A. (2003). Conservation of Endo16 expression in sea urchins despite evolutionary divergence in both cis and trans-acting components of transcriptional regulation. *Development* **130**, 4187-4199.
Royo, J. L., Hidalgo, C., Roncero, Y., Seda, M. A., Akalin, A., Lenhard, B., Casares, F. and Gómez-Skarmeta, J. L. (2011). Dissecting the transcriptional regulatory properties of human chromosome 16 highly conserved non-coding regions. *PLoS ONE* **6**, e24824.
Sandelin, A., Bailey, P., Bruce, S., Engström, P. G., Klos, J. M., Wasserman, W. W., Ericson, J. and Lenhard, B. (2004). Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics* **5**, 99.
Sandmann, T., Jensen, L. J., Jakobsen, J. S., Karzynski, M. M., Eichenlaub, M. P., Bork, P. and Furlong, E. E. (2006). A temporal map of transcription factor activity: mef2 directly regulates target genes at all stages of muscle development. *Dev. Cell* **10**, 797-807.
Sato, S., Ikeda, K., Shioi, G., Nakao, K., Yajima, H. and Kawakami, K. (2012). Regulation of Six1 expression by evolutionarily conserved enhancers in tetrapods. *Dev. Biol.* **368**, 95-108.
Schmidt, D., Wilson, M. D., Ballester, B., Schwalie, P. C., Brown, G. D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C. P., Mackay, S. et al. (2010). Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* **328**, 1036-1040.
Schmidt, S. F., Jørgensen, M., Chen, Y., Nielsen, R., Sandelin, A. and Mandrup, S. (2011). Cross species comparison of C/EBPα and PPARγ profiles in mouse and human adipocytes reveals interdependent retention of binding sites. *BMC Genomics* **12**, 152.
Senger, K., Armstrong, G. W., Rowell, W. J., Kwan, J. M., Markstein, M. and Levine, M. (2004). Immunity regulatory DNAs share common organizational features in Drosophila. *Mol. Cell* **13**, 19-32.
Shin, J. T., Priest, J. R., Ovcharenko, I., Ronco, A., Moore, R. K., Burns, C. G. and MacRae, C. A. (2005). Human-zebrafish non-coding conserved elements act in vivo to regulate transcription. *Nucleic Acids Res.* **33**, 5437-5445.
Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S. et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034-1050.
Song, L., Zhang, Z., Grasfeder, L. L., Boyle, A. P., Giresi, P. G., Lee, B. K., Sheffield, N. C., Gräf, S., Huss, M., Keefe, D. et al. (2011). Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. *Genome Res.* **21**, 1757-1767.
Sudou, N., Yamamoto, S., Ogino, H. and Taira, M. (2012). Dynamic in vivo binding of transcription factors to cis-regulatory modules of cer and gsc in the stepwise formation of the Spemann-Mangold organizer. *Development* **139**, 1651-1661.
Taher, L., McGaughey, D. M., Maragh, S., Aneas, I., Bessling, S. L., Miller, W., Nobrega, M. A., McCallion, A. S. and Ovcharenko, I. (2011). Genome-wide identification of conserved regulatory function in diverged sequences. *Genome Res.* **21**, 1139-1149.
Takahashi, H., Mitani, Y., Satoh, G. and Satoh, N. (1999). Evolutionary alterations of the minimal promoter for notochord-specific Brachyury expression in ascidian embryos. *Development* **126**, 3725-3734.
Teng, Y., Girard, L., Ferreira, H. B., Sternberg, P. W. and Emmons, S. W. (2004). Dissection of cis-regulatory elements in the C. elegans Hox gene egl-5 promoter. *Dev. Biol.* **276**, 476-492.
Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B. et al. (2012). The accessible chromatin landscape of the human genome. *Nature* **489**, 75-82.
Vavouri, T. and Lehner, B. (2009). Conserved noncoding elements and the evolution of animal body plans. *BioEssays.* **31**, 727-735.
Vavouri, T., McEwen, G. K., Woolfe, A., Gilks, W. R. and Elgar, G. (2006). Defining a genomic radius for long-range enhancer action: duplicated conserved non-coding elements hold the key. *Trends Genet.* **22**, 5-10.
Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A. et al. (2001). The sequence of the human genome. *Science* **291**, 1304-1351.
Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F. et al. (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature* **457**, 854-858.

Wilczyński, B. and Furlong, E. E. (2010). Dynamic CRM occupancy reflects a temporal map of developmental progression. *Mol. Syst. Biol.* **6**, 383.
Wilson, M. D., Barbosa-Morais, N. L., Schmidt, D., Conboy, C. M., Vanes, L., Tybulewicz, V. L., Fisher, E. M., Tavaré, S. and Odom, D. T. (2008). Species-specific transcription in mice carrying human chromosome 21. *Science* **322**, 434-438.
Wittkopp, P. J. and Kalay, G. (2012). Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat. Rev. Genet.* **13**, 59-69.
Woolfe, A., Goodson, M., Goode, D. K., Snell, P., McEwen, G. K., Vavouri, T., Smith, S. F., North, P., Callaway, H., Kelly, K. et al. (2005). Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* **3**, e7.
Wray, G. A. (2007). The evolutionary significance of cis-regulatory mutations. *Nat. Rev. Genet.* **8**, 206-216.
Wray, G. A., Hahn, M. W., Abouheif, E., Balhoff, J. P., Pizer, M., Rockman, M. V. and Romano, L. A. (2003). The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.* **20**, 1377-1419.