12 Organization of Conserved Elements Near Key Developmental Regulators in Vertebrate Genomes

Adam Woolfe[1] and Greg Elgar

School of Biological and Chemical Sciences, Queen Mary, University of London, London E1 4NS, United Kingdom

[1] Current address: Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, Rockville, Maryland 20870

Advances in Genetics, Vol. 61
0065-2660/08 $35.00
DOI: 10.1016/S0065-2660(07)00012-0

ABSTRACT

Sequence conservation has traditionally been used as a means to target functional regions of complex genomes. In addition to its use in identifying coding regions of genes, the recent availability of whole genome data for a number of vertebrates has permitted high-resolution analyses of the noncoding "dark matter" of the genome. This has resulted in the identification of a large number of highly conserved sequence elements that appear to be preserved in all bony vertebrates. Further positional analysis of these conserved noncoding elements (CNEs) in the genome demonstrates that they cluster around genes involved in developmental regulation. This chapter describes the identification and characterization of these elements, with particular reference to their composition and organization.

I. INTRODUCTION

Complex multicellular organisms must overcome the fact that whil all constituent cells contain the same genetic instructions, they do not all express the same set of genes. Specific populations or groups of cells must be able to differentiate, specialize, and perform defined roles, so enabling the survival of the whole. Once a differentiated state has become established, further changes in gene expression allow the organism to respond to insult or challenge. The most complex, dynamic, and exquisitely controlled period of an organism's life occurs during development, from single cell to fully differentiated state. In vertebrates, this process results in billions of differentiated cells, communicating and functioning in a particular and coordinated manner, responding and adapting as organized cohorts.

Biologists have been studying how genes are regulated for many decades and despite, or perhaps because of, its complex nature, a great deal of attention has been focused on the control of vertebrate development. This has resulted in the elucidation of many of the regulatory genes and pathways involved, and yet surprisingly, there is scant knowledge of either the molecular mechanisms or the underlying regulatory language involved.

Clearly, one of the limitations in identifying regulatory elements around vertebrate genes was, until a few years ago, the lack of genomic sequence data available. The completion of the human genome in 2001, followed by those of a host of other vertebrates in the proceeding few years, provided an opportunity to mine these vast datasets for specific sequences that might have function. The simplest and most commonly used approach has been to identify regions of noncoding DNA that have been conserved between two or more species, implying that some functional constraint has prevented these sequences from mutating at a neutral rate. However, despite the simplicity of the approach, interpretation of the results has often been confounded by a number of factors: rates of evolution are not homogenous across or between genomes, regulatory sequences are often poorly conserved at the primary sequence level, and our knowledge of the regulatory language of the genome is rather limited. In fact, regulatory sequences, it appears, are often not conserved at all.

However, under certain circumstances, some of these analyses have been insightful, particularly those aimed at identifying sequences with regulatory potential around genes that coordinate early development. Given the critical nature and exquisite complexity of developmental processes, both in terms of time and in terms of 3-dimensional space, one might expect to find an abundance of regulatory control elements encoded in the genomic DNA surrounding these genes. This chapter describes the identification of such a set of sequences, and how these sequences are organized with respect to the genes upon which they act in vertebrate genomes.

II. GENE-REGULATORY NETWORKS IN DEVELOPMENT

Animal development is thought to be controlled by progression through a number of transcriptional states that are transiently positioned in embryonic space (Levine and Davidson, 2005). This is orchestrated by regulated expression of transcription factor genes in a specific spatiotemporal manner. These genes are controlled by multiple *cis*-regulatory modules (CRMs) (often termed “enhancers,” “repressors,” or “silencers”) that act as sites for combinatorial protein binding in order to actuate complex patterns of expression. This complex set of interactions during development can be represented in the form of a gene regulatory network (GRN). GRNs involving specific developmental pathways have so far been dissected in a number of invertebrates. These include endomesoderm specification in the sea urchin (Davidson *et al.*, 2002), dorsal-ventral specification in *Drosophila melanogaster* (Levine and Davidson, 2005; Ochoa-Espinosa *et al.*, 2005), and the general specification of developmental pathways in *Ciona intestinalis* (Shi *et al.*, 2005). Evidence on which these GRNs are based derives from large-scale experimental efforts which involve obtaining spatial and temporal expression data on all genes in the network and the use of large-scale perturbation analysis to see the effect of gene deletion on the expression of other genes in the network (Davidson *et al.*, 2002). Among the more complex vertebrates, *Xenopus laevis* has been proposed as a model system in which to study GRNs, because of the ease with which genes can be manipulated, and a small GRN involving mesoderm specification has already been proposed (Koide *et al.*, 2005; Loose and Patient, 2004). However, the problem of identifying the CRMs controlling these networks is still a major challenge.

III. IDENTIFICATION OF EVOLUTIONARILY CONSTRAINED SEQUENCES USING PHYLOGENETIC FOOTPRINTING

Prior to the availability of whole genome sequences, classical searches for distal *cis*-regulatory elements typically involved various trial-and-error strategies. Experimental approaches to the identification of regulatory elements include deletion constructs of upstream sequences to determine the minimal sequences necessary for transcription in cell culture-based systems, DNAse I hypersensitivity studies to identify sequences potentially available for transcription factor binding, and DNA footprinting to determine sequences that bind various regulatory proteins (see Pennacchio and Rubin, 2001). Large-scale promoter- and enhancer-trapping studies have also been carried out in mice (Durick *et al.*, 1999) and zebra fish (Ellingsen *et al.*, 2005), but on the whole, laboratory attempts at identification and characterization of distal regulatory elements have been unguided, highly laborious, and time-consuming. Access to the sequences of several vertebrate genomes has presented an unprecedented opportunity for the discovery of functional elements in the human genome through comparative genomics. The discovery of putative functional elements through comparison of sequences from several species, known as phylogenetic footprinting, is based on the assumption that these elements evolve more slowly than surrounding nonfunctional DNA, as they are under negative (purifying) selection. Thus, sequences that are more highly conserved than would be expected under a reasonable model of neutral evolution are likely to be important for function. One of the key decisions inherent in phylogenetic footprinting is the choice of organisms with which the comparison will be made. Given the common and fundamental nature of early embryogenesis across the vertebrate lineage, it might be expected that many of the regulatory instructions hardwired into the genome would be highly conserved across large evolutionary distances.

IV. SEARCHES FOR REGULATORY ELEMENTS USING EVOLUTIONARY CONSERVATION

Genomic comparisons have frequently been used as a method to identify regulatory elements. Early studies using comparisons between distantly related *D. melanogaster* species proved successful in identifying conserved enhancer elements (Martinez-cruzado *et al.*, 1988), as the rapid rate of evolution in this species helped resolve functional sequences from background conservation. The release of a number of draft vertebrate genomes has in parallel spurred the development of appropriate genomic alignment, visualization, and analytical bioinformatics tools have made large-scale sequence comparisons not only possible but an increasingly popular approach for the discovery of functional elements in these genomes. Several studies have attempted to identify regulatory elements in mammals using human–mouse comparisons (e.g., Frazer *et al.*, 2004; Göttgens *et al.*, 2000; Hardison *et al.*, 1997; Oeltjen *et al.*, 1997; Wasserman *et al.*, 2000). Unfortunately, owing to differences in mutation rates across the genome, relatively small evolutionary divergence between mammals and the slow rate of neutral divergence among vertebrates, many more sequences may be conserved than actually play functional roles (Tautz, 2000). Indeed, although ~40% of the human and mouse genomes are alignable, only ~5% is estimated to be under evolutionary constraint (Waterston *et al.*, 2002). Consequently, specific conservation criteria were proposed to distinguish functional conservation from background. An arbitrary criterion of 70% identity over at least 100 bp of ungapped alignment (which is above the average rate of neutral conservation) between human and mouse sequences has been used to successfully identify a number of regulatory elements (Göttgens *et al.*, 2000; Loots *et al.*, 2000). Using this criterion across whole-genome human–mouse alignments identified ~327,000 conserved elements, making up around 1% of the human genome, which were located in noncoding regions and had little or no evidence of transcription (Dermitzakis *et al.*, 2003, 2005). These sequences appear to be distributed uniformly across the genome and are negatively correlated with the distribution of genes, suggesting roles which are distance-independent or that are not directly involved in gene expression. An alternative approach for the identification of noncoding constrained elements was proposed by Margulies *et al.* (2003), who devised two strategies based on parsimony and binomial-based models applied to multispecies alignments. Unlike simple percent-identity-based approaches, these models take into account the derived local neutral mutation rate as well as the divergence times between sequences based on a phylogenetic tree. Combining these approaches, on a 1.7 Mb region around the *CFTR* locus, they were able to successfully distinguish between neutrally evolving sequence such as known ancestral repeats and constrained elements such as exons, and identified a large number of conserved elements, ~70% of which were located in noncoding regions. Many more constrained sequences were identified using this multiple-alignment approach than could be identified using human–mouse pairwise alignments alone, demonstrating the power of multispecies alignments (Margulies *et al.*, 2006).

Nevertheless, it is currently unclear what proportion of these constrained noncoding elements is regulatory or even functional at all. This was highlighted by Nobrega *et al.* (2004) who deleted two large noncoding regions of roughly 1 Mb in size on mouse chromosomes 3 and 19. These regions contain a total of 1243 sequences that are >70% identical across at least 100 bp between human and mouse (a small number of which were also conserved in chicken and frog) and yet their deletion caused no loss of viability or any other overt phenotypic changes. Quantitative polymerase chain reaction (PCR) analysis showed that expression levels of the genes flanking the deleted sections were unaffected in all but 2 of the 108 tissues assayed. In addition, enhancer assays in transgenic mice of 15 of the most highly conserved elements in these regions (10 of which were conserved in chicken and 5 in frog) found only one that upregulated the reporter gene in a tissue-specific manner. Although it is possible that these deletions may cause abnormalities undetected in this time setting or environmental context, this study highlights the possibility that many sequences that have remained conserved across large evolutionary distances may not play critical functional roles. Furthermore, the ability to use comparative genome alignments to discover thousands of potentially functional regulatory elements in mammalian genomes overwhelms our current ability to test their function *in vivo*.

Commonly used techniques for testing putative distal regulatory sequences for enhancer activity, such as transgenic reporter gene assays in developing mouse embryos, are still very expensive, time-consuming, and laborious. Cheaper and faster *in vivo* reporter gene assays have been developed for use in other model organisms such as frog, zebra fish, and medaka (Müller *et al.*, 2002), although it would still take many decades to test even a fraction of the elements thought to be functional in mammals. In light of the inability of current technologies to test large numbers of mammalian-conserved elements with "regulatory potential," it is important to be able to prioritize a smaller set of elements with high regulatory potential for more focused studies. To address this, Bejerano *et al.* (2004) searched for sequences that were identical over at least 200 bp between human, mouse, and rat (termed "ultraconserved" elements) and identified 481 such sequences, of which over half were located in noncoding regions. By contrast with other mammalian-conserved sequences, these elements often clustered in the vicinity of genes involved in transcriptional regulation and/or development and overlapped a number of known enhancers, suggesting that they are likely to function as CRMs. Another highly successful approach for prioritizing CRMs is the identification of sequences that are conserved across extreme evolutionary distances such as fish and mammals. Teleost fish are well suited for comparisons with mammals, as they last shared a common ancestor 400–450 million years ago and therefore it is assumed that only critical functional sequences would be conserved between genomes that are otherwise diverged. One teleost genome, in particular, that of the puffer fish *Takifugu rubripes*, has been used extensively for the discovery of CRMs.

V. *TAKIFUGU RUBRIPES*: A COMPACT MODEL GENOME

The sequencing of the Japanese puffer fish *Takifugu rubripes* (commonly known as *Fugu*) was first proposed by Sydney Brenner and colleagues in 1993 as a compact model vertebrate genome (Brenner *et al.*, 1993). They showed *Fugu* has one of the smallest genomes of any known vertebrate at around 390 Mb in length (around one eighth the size of the human genome), but as a vertebrate, it has a similar complement of genes to that of mammals (Elgar *et al.*, 1996). In addition, given the enormous cost of sequencing large genomes at that point, *Fugu* represented a genome that could be completed within a few years at a fraction of the cost of the human genome whil providing a resource for its annotation. The publication of the draft version of the genome in 2002 (Aparicio *et al.*, 2002) hailed the release of only the second vertebrate genome to be sequenced after human. The justification for sequencing was demonstrated immediately, with the first comparison of the *Fugu* and human genome revealing more than 1000 genes that had previously remained unidentified (Aparicio *et al.*, 2002). The compact nature of the *Fugu* genome is, principally, due to the low abundance of repeat sequences and a reduction in the size of intronic and intergenic sequences, with a resultant increase in gene density. The remaining genomic sequence is, consequently, enriched for regulatory and other functional elements, helping reduce the search space for such elements in comparative analyses with other genomes. Several other interesting aspects of fish in general have spurred the sequencing of a number of other teleosts. Teleosts are a highly diverse group, making up half of all extant vertebrate species (Nelson, 1994). Teleosts underwent a whole genome-duplication event around 330 million years ago that coincided with this huge burst in diversification (Hoegg *et al.*, 2004) leading many to believe that the two events were linked. Some teleosts are used as experimental developmental models, in particular zebra fish and medaka. Fish are therefore excellent models in the study of speciation, genome evolution, gene duplication, and development.

VI. IDENTIFICATION OF ENHANCER ELEMENTS THROUGH FISH-MAMMAL COMPARISONS

Although initial interest in the *Fugu* genome centered on its application for gene identification, several studies preceding the release of the genome sequence indicated that fish–mammal comparisons may also be useful in the identification of some regulatory enhancers. The first example of conserved noncoding sequences (CNSs) identified between *Fugu* and mammals was found around the *Hoxb-1* gene and were shown to be able to recapitulate *Hoxb-1* neuroectoderm expression in developing mouse embryos (Marshall *et al.*, 1994). This was followed up by a similar survey around the *Hoxb4* gene which found a number of other conserved enhancer elements through transgenic testing in mouse (Aparicio *et al.*, 1995). More recently, comparative analyses with *Fugu* identified a global control region responsible for tissue-specific expression of the HoxD cluster by pinpointing a core sequence within a 40-kb region known to harbor this element (Spitz *et al.*, 2003). Several other studies utilizing *Fugu*–mammal comparisons have identified CNSs indicative of important regulatory elements and in each case the genes under study are implicated in developmental control. Such studies have identified enhancers responsible for the regulation of *PAX6* (Griffin *et al.*, 2002; Kammandel *et al.*, 1999; Miles *et al.*, 1998), *SOX9* (Bagheri-Fam *et al.*, 2001), the Dlx genes (Ghanem *et al.*, 2003), *Wnt-1* (Rowitch *et al.*, 1998), *PAX9/Nkx2–9* (Santagati *et al.*, 2001, 2003), and the Iroquois clusters (De la Calle-Mustienes *et al.*, 2005). In a pioneering study aimed at characterizing intriguing regions of low gene density in the human genome, known as gene deserts, Nobrega *et al.* (2003) identified over 1000 conserved elements between human and mouse around and within the introns of *DACH1*, a gene involved in development of the brain, limbs, and sensory organs. To narrow the search for sequences likely to be functional, they looked for sequences also conserved in Frog and *Fugu*, shortening the list to 32 candidates. Nine of these were tested in reporter gene assays in mouse embryos and seven were found to reproducibly drive β-galactosidase expression in a manner that recapitulated several aspects of *DACH1* endogenous expression. These studies demonstrate the power of using highly divergent sequences to prioritize sequences likely to be *cis*-regulatory in function. A growing number of acronym-based names, such as CNSs and multi-species conserved sequences (MCS), have been used to describe noncoding sequences under evolutionary constraint (Table 12.1), reflective of the varied nature of the identification processes (i.e., different neutral models and across different evolutionary distances) as well as current ambiguity in relation to their function. For ease of reference, I have chosen the acronym conserved noncoding elements (CNEs) to refer specifically to a noncoding sequence which is conserved between fish and mammals and therefore has a high regulatory potential. Despite the identification of a number of *Fugu*–mammal CNSs within a limited

Table 12.1. Acronyms Used for Conserved Noncoding Regions in Vertebrate Genomes

Acronym	Meaning	Reference
ANCOR	Ancestral noncoding conserved region	Aloni and Lancet, 2005
CNC	Conserved noncoding	Couronne *et al.*, 2003
CNE	**Conserved noncoding element**	**Woolfe *et al.*, 2005**
CNG	Conserved nongenic	Dermitzakis *et al.*, 2003
CNS	Conserved noncoding sequence	Dubchak *et al.*, 2000
CST	Conserved sequence tag	Mignone *et al.*, 2003
ECR	Evolutionary conserved region	Ovcharenko *et al.*, 2004
HCR	Highly conserved region	Duret and Bucher, 1997
MCS	Multispecies conserved sequence	Thomas *et al.*, 2003
UCE	Ultraconserved element	Bejerano *et al.*, 2004

Notes: References refer to the chapter in which the acronym was first used. In this study, the acronym CNE (highlighted in bold) is used to refer specifically to DNA sequences conserved from mammals to fish. Table adapted from Aloni and Lancet, 2005.

number of well-studied developmental genes, it was not clear whether this was indicative of a more extensive genome-wide trend and whether similar CNEs were located around other types of genes. To answer this question, a pairwise comparison of the *Fugu* and human genomes was performed to identify genome-wide conservation of noncoding sequences (Woolfe *et al.*, 2005).

VII. FISH-MAMMAL CONSERVED NONCODING ELEMENTS ARE ASSOCIATED WITH VERTEBRATE DEVELOPMENT

The *Fugu* genome was masked for the majority of coding and tRNA content and the remaining regions compared to the human genome using basic local alignment search tool (MegaBLAST) with a stringent "seeding" word size of 20 bp. All resulting matches were filtered to include alignments of over 100 bp in length and exclude any repetitive elements, protein coding or ncRNA sequences, that had been missed. One thousand three hunhdred and seventy-three CNEs were identified in this way, with little or no evidence of transcription [except for ~6% located within untranslated regions (UTRs) of known mRNA molecules]. Unsurprisingly, the majority of the CNEs are also conserved in other mammals such as the mouse and rat as well as in the chicken and zebra fish genomes, indicating these sequences are likely to be common to all bony vertebrates. By contrast, no significant similarity with these CNEs can be found within invertebrate chordate or other invertebrate genomes, suggesting these sequences are a vertebrate innovation. In human, CNEs are found in all chromosomes except chromosome 21 and Y, but their distribution is highly clustered. One hundred and sixty-five clusters were defined, with over 85% of clusters containing five or more CNEs. A statistical analysis of the gene ontology (GO) assignments and InterPro structural domains for all genes located within these clusters found 12 of the 13 most overrepresented GO terms relate to transcriptional regulation and development (Table 12.2) as well as enrichment for DNA-binding motifs (such as the C_2H_2 zinc finger domain), in particular the homeobox domain. Indeed, over 96% of clusters are located in the vicinity of one or more genes with such functional assignments. Initial observations of tight association of CNEs with genes involved in transcriptional regulation and/or development (which for ease are referred to as *trans-dev* genes) were therefore confirmed on a genome-wide scale.

A number of the *trans-dev* genes identified have previously been shown to have highly conserved *cis*-regulatory elements associated with them. The association of noncoding sequences conserved deep in the vertebrate lineage with *trans-dev* genes has since been confirmed by a number of other studies (Ahituv *et al.*, 2005; Ovcharenko *et al.*, 2005; Sandelin *et al.*, 2004; Sironi *et al.*, 2005). CNEs tend to be located in regions of low gene density, also referred to as "gene deserts" (Nobrega *et al.*, 2003; Ovcharenko *et al.*, 2005), confirming the likelihood that these regions harbor large numbers of distal CRMs. Sensitive

Table 12.2. Table of Overrepresented GO Terms in 170 Annotated Genes with *Fugu*–Human Conservation in the UTR of both Orthologues

GO Id	GO term	Group Count	Total Count	*P*-value
GO: 0003700	Transcription factor activity	40	1292	4.92×10^{-26}
GO: 0045449	Regulation of transcription	60	2922	4.71×10^{-21}
GO: 0019219	Nucleobase, nucleoside, nucleotide, and nucleic acid metabolism	60	2953	8.14×10^{-21}
GO: 0031323	Regulation of cellular metabolism	60	2991	1.9×10^{-20}
GO: 0006355	Regulation of transcription, DNA-dependent	57	2799	6.07×10^{-20}
GO:0051244	Regulation of cellular physiological process	70	3907	6.07×10^{-20}
GO: 0019222	Regulation of metabolism	61	3134	6.18×10^{-20}
GO: 0006350	Transcription	60	3061	7.16×10^{-20}
GO: 0006351	Transcription, DNA-dependent	57	2875	3.58×10^{-19}
GO: 0005634	Nucleus	77	5243	$8.45{\times}10^{-15}$
GO: 0007399	Neurogenesis	17	448	2.62×10^{-07}
GO: 0043227	Membrane-bound organelle	80	7195	3.71×10^{-07}
GO: 0043231	Intracellular membrane-bound organelle	80	7195	3.71×10^{-07}
GO:0007492	Endoderm development	2	2	0.00157
GO:0007420	Brain development	4	37	0.00332
GO:0048523	Negative regulation of cellular process	13	592	0.00401

Notes: Group count indicates the number of genes from within the 170 that posesss a particular GO term, whereas Total count refers to the number of genes in the human genome with that GO term. The *P*-value is an indicator as to how enriched the group count is compared to the total count.

multiple alignments using additional mammalian orthologous sequence around a number of the CNE cluster regions identified more than double the number of CNEs of over 100 bp in length, demonstrating the power of this approach. To confirm their regulatory potential, 25 CNEs, derived from both whole genome comparison and multiple-alignment approaches, located in clusters surrounding four different *trans-dev* genes (*SOX21*, *HLXB9*, *SHH*, and *PAX6*), were tested for enhancer activity using a transient enhancer green fluorescent protein (GFP) reporter-gene assay in zebrafish (*Danio rerio*) embryos. Twenty-three of the 25 CNEs tested were shown to have reproducible enhancer activity, inducing expression of a GFP reporter gene in a temporal and tissue-specific manner that frequently coincides with the endogenous expression domains of their nearby *trans-dev* gene. As this assay measures only upregulation of gene expression, those CNEs that showed no enhancer activity may represent negative regulatory elements such as silencers or insulators, or may be involved in indirect processes such as chromatin remodeling. These results demonstrate the frequent functional nature of CNEs identified through fish–mammal comparisons. The extreme conservation of these CNEs across the vertebrate lineage together with their tight association with intricately regulated early developmental genes implies that they are enriched for putative CRMs.

VIII. HIGH-RESOLUTION ANALYSIS OF THE ORGANIZATION OF CNEs AROUND KEY DEVELOPMENTAL REGULATORS

In order to examine the conserved regulatory architecture around key developmental genes in greater detail, multiple alignments have been performed using MLAGAN (Brudno *et al.*, 2003a,b), and the resulting data stored in a publicly accessible relational database, *co*nserved *no*n-co*d*ing *o*rthologous *r*egions (CONDOR), that also includes functional data on selected CNEs [Woolfe *et al.* (submitted for publication)] and http://condor.fugu.biology.qmul.ac.uk/). Multiple alignments allow greater sensitivity, as the probability that a CNE will occur in more than two genomes around the same gene by chance is greatly reduced compared with a pairwise alignment. Despite evidence that there is a degree of shuffling of conserved regulatory signatures (Sanges *et al.*, 2006), the vast majority of CNEs identified associated with *trans-dev* genes appear to be co-linear (Sun *et al.*, 2006) and are thus well suited to this form of analysis, which uses "chained anchors" along the sequence, that is, small blocks of well-conserved sequence along the alignment. Another aligner, shuffle LAGAN (SLAGAN) (Brudno *et al.*, 2003c), can identify shuffled CNEs.

CONDOR contains detailed data on 7000 CNEs spanning nearly 100 gene regions and provides an unprecedented opportunity to look at the distribution of CNEs on a large scale.

IX. GENERAL GENOMIC ENVIRONMENT AROUND CNEs

CNEs are by definition located in noncoding sequence, although the genic environment surrounding the noncoding sequence can be very different. The distribution of CNEs around *trans-dev* genes is of particular note because it contradicts the traditional view that *cis*-regulatory sequences are located just upstream of the promoter. Figure 12.1 shows the proportion of CNEs from CONDOR that are located in introns (of either the *trans-dev* gene upon which the CNE acts, or of neighboring genes), between genes, or in UTRs. Around 4% of CNEs appear to be located in UTRs, and this is discussed in Section X. Interestingly, nearly a third of CNEs are found in introns, both of the *trans-dev* gene upon which they act (18%) and in the introns of neighboring genes (12%). When CNEs are located in neighboring genes, these genes are invariably found in conserved synteny in all vertebrates, presumably as a result of having to maintain the regulatory repertoire of the *trans-dev* gene intact. By examining the position of 4950 CNEs that are in the vicinity of just a single *trans-dev* gene, we can determine whether there is any positional bias toward maintaining either upstream or downstream CNEs. Figure 12.2 demonstrates that there are almost as

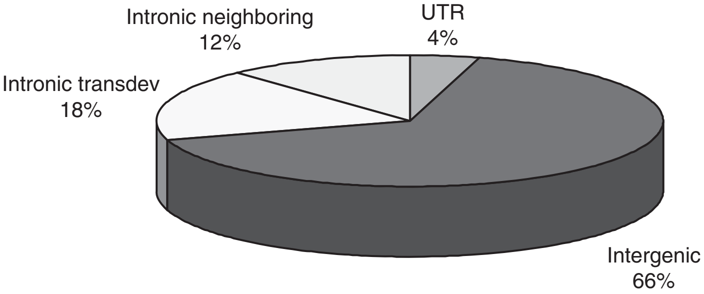

Figure 12.1. Distribution of CNEs across *trans-dev* genes. CNEs are classified as being within untranslated regions (UTR), between genes (Intergenic), within the introns of a *trans-dev* gene (Intronic *trans-dev*), or within the introns of a neighboring (non-*trans-dev*) gene (Intronic neighboring).

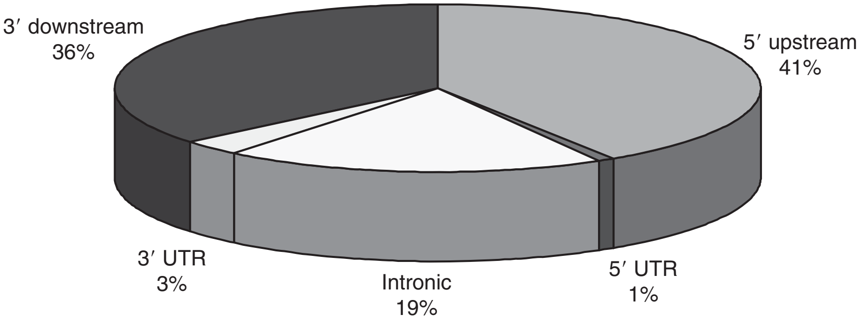

Figure 12.2. Position of CNEs relative to the *trans-dev* gene they are associated with.

many CNEs located downstream of *trans-dev* genes as there are upstream. Indeed, if we include intronic CNEs as well as those located in the 3′ UTR, over 58% of CNEs are located 3′ of the transcription start site (TSS).

Not surprisingly, the distribution of CNEs across any one specific *trans-dev* gene varies considerably, but positional bias of CNEs either 5′ or 3′ generally correlates with a similar bias in the amount of flanking noncoding sequence at that position. For example, the largest bias in both CNE position and flanking DNA is seen around the human *BCL11A* gene, which has virtually no flanking DNA 5′ of the gene (and no CNEs), whil 80% of CNEs are located within a large (2.21 Mb) gene desert located 3′ of the gene. Conversely, 92% of the CNEs around the *FOXD3* gene are found in a 450-kb gene desert 5′ of the gene, whereas the region 3′ is gene rich and has no CNEs (Fig. 12.3). The variable distribution of CNEs therefore suggests that they can function irrespective of position around their associated gene unlike other position-specific *cis*-regulatory elements such as promoters.

X. CNEs PRESENT IN TRANSCRIPTS

Although all CNEs that are located within the introns of genes are by definition transcribed, only ~5% of CNEs are located in UTRs and therefore constitute part of the mature mRNA molecule. These are dealt with in more detail below. To investigate whether the remaining 95% "nongenic" CNEs are likely to form part of mature transcripts, sequences were BLAST searched against expressed sequence tags (ESTs) databases encompassing all currently deposited ESTs from vertebrates. Seventy-seven percent of these CNEs have no significant hits to ESTs, 11.1% have one hit, 7.2% have between two and four hits to ESTs, and 4.7% have five or more hits. In contrast, BLAST searches of 1000 randomly conserved elements overlapping known coding exons from CONDOR regions identified just 4.5% with two to four hits to ESTs, and 95.5% with five or more, demonstrating that transcribed sequences are characterized by hits to large numbers of ESTs. The vast majority of CNEs therefore have little or no evidence of transcription, indicating likely roles as CRMs. The small proportion of CNEs with larger hits to ESTs may represent unannotated coding sequences or noncoding RNAs, but equally likely, these sequences may represent good candidates for CRMs. A number of these sequences (e.g., CRCNE00007308 located in the gene desert between *IRX3* and *IRX5* and CRCNE00006651 located in the gene desert between *IRX1* and *IRX2*, both of which have five hits to ESTs) have been shown to drive reporter constructs in zebra fish embryos in a spatial- and temporal-specific manner (De la Calle-Mustienes *et al.*, 2005), suggesting at least some of these sequences are regulatory in function. This observation also indicates the possibility that some regulatory elements are part of sequences that

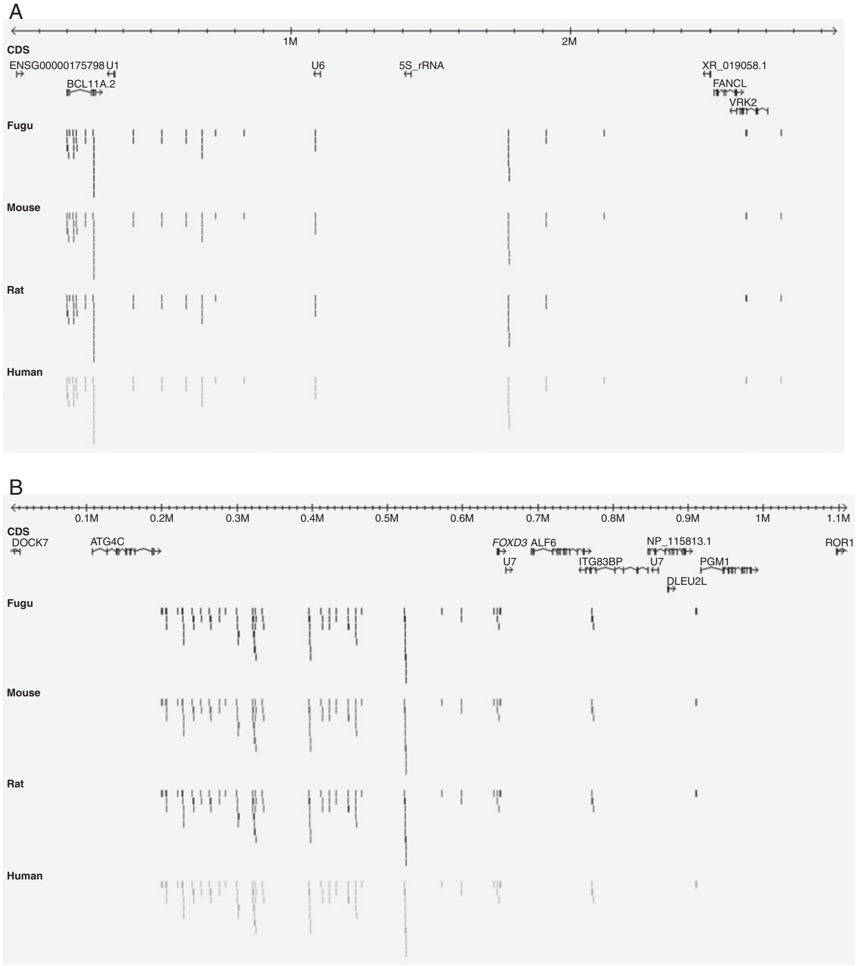

Figure 12.3. CONDOR view of (A) *BCL11A* and (B) *FOXD3* gene regions showing positions of CNEs in human, rat, mouse, and Fugu. Note that the majority of the CNEs are 3′ of the first exon of the *BCL11A* gene, and 5′ of the *FOXD3* gene, correlating with the location of gene deserts in each region.

play a dual function as coding sequences or work through an RNA intermediate (e.g., Jones and Flavell, 2005) in addition to or as part of their regulatory function. There is also a possibility that the matching ESTs derive from genomic contamination and are not transcribed at all or they are the product of illegitimate transcription (Sorek and Safer, 2003).

XI. CNEs LOCATED WITHIN UTRs

A small number of CNEs are located within the UTRs of some mammalian *trans-dev*, or neighboring genes (presence within a UTR of a gene in *Fugu* is not currently discernable due to lack of full-length cDNA data). Within the entire CNE dataset (around all *trans-dev* genes in CONDOR), 279 sequences (4.2%) are located in a UTR. Of these, 85% are located in the UTR of a gene annotated as *trans-dev* and the rest are located in UTRs of neighboring (non-*trans-dev*) genes. Over 74% of CNEs located in *trans-dev* UTRs are located in the 3′ UTR, reflecting the increased size of 3′ UTRs compared to 5′ UTRs in vertebrate genomes (the average length of 5′ UTRs in human Ensembl is only 255 bp, compared with 989 bp for 3′ UTRs). By contrast, equal numbers of CNEs are located in the 5′ and 3′ UTRs of genes neighboring *trans-dev* genes. As these sequences, unlike other CNEs, are transcribed, it is not known whether they function at the pre-transcriptional (i.e. *cis*-regulatory) or post-transcriptional level (e.g., directing mRNA stability).

Nevertheless, a genome-wide analysis between all human UTRs and the *Fugu* genome indicates that sequence conservation is rare, and in most cases coincides with a *trans-dev* gene region that already has catalogued CNEs. Furthermore, there are at least eight examples where a CNE is located in the UTR of a human gene for which there is either no (gene) orthologue in Fugu, or the orthologue is in a different location to the CNE in Fugu. This sheds light on the possible function of CNEs in UTRs and the utility of using highly diverged genomes in such cases. Five examples have no orthologue in the *Fugu* genome, with three of these being primate-specific genes, indicating that the CNE has been coincidentally incorporated into the UTR of a "new" gene sometime in primate evolution. The other two are located in UTRs of the *HOXA7* and *HOXB7* genes in human, genes that have been lost sometime during the evolution of Fugu (Aparicio *et al.*, 1997), yet the CNEs have been retained in the same place, located between *Hoxa/b6* and *Hoxa/b8*, indicating a putative *cis*-regulatory role for the surrounding *Hox* genes rather than post-transcriptional regulation of *Hoxa7/b7*. The final three examples also suggest that CNEs are unlikely to be involved in post-transcriptional regulation of the genes in which they are located. These are located in *Fugu* regions where gene synteny around the CNE has been retained, but the gene in which they are located in human is present elsewhere in the *Fugu* genome. An example of this can be seen in Fig. 12.4. A CNE (corresponding to CNE CRCNE00007244 in CONDOR) is located within the UTR of an ancient, relatively uncharacterized gene *Q9Y2K8*, which is conserved from mammals through to *Caenorhabditis elegans* (Ensembl Compara v.36). It is located upstream of the Iroquois B cluster (made up of Irx3, -5, and -6), a region characterized by large numbers of CNEs which extends through the neighboring genes (including *Q9Y2K8)* across a region encompassing almost 4.9 Mb on human chromosome 16. CRCNE00007244 and *Q9Y2K8*

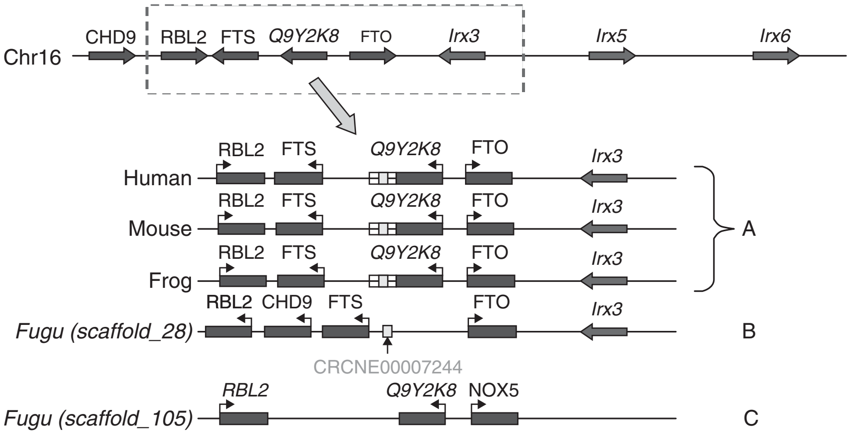

Figure 12.4. Location of CNEs in the UTR of nondevelopmental genes is unlikely to play a role in their function. CNE CRCNE00007244 (shaded box) is located in *IRX3, -5, and -6* region and is found in the UTR of a relatively uncharacterized gene *Q9Y2K8* (*KIAA1005*). This gene is conserved back to *C. elegans* and the gene order is preserved in this region in tetrapods from mammals down to frog (A), indicating the likely ancestral state. In *Fugu*, this gene is no longer present in the *Irx3, -5, and -6* region (scaffold_28) (B), but the CNE remains, indicating it plays no role in this gene and is associated with neighboring genes, or the IRX cluster. The synteny of the rest of the genes in this region has remained the same in *Fugu*, although RBL2 and CHD9 have undergone an inversion event, so they are in the opposite orientation relative to their orthologues in tetrapods. The orthologue of *Q9Y2K8* is present in *Fugu* but is located on scaffold_105 (C). Presence of a second *RBL2* gene downstream of *Q9Y2K8* on this scaffold suggests these genes underwent a fish-specific duplication and the copy on scaffold_28 was lost through nonfunctionalizsation over evolution.

are conserved in the same relative position in all currently available tetrapod genomes as is the gene order across the rest of the IrxB cluster. In *Fugu*, the CNE is located in the same relative position (on scaffold_28, ensembl) containing the IrxB cluster, but *Q9Y2K8* is located on another scaffold. Interestingly, in *Fugu*, *Q9Y2K8* is positioned next to another gene paralogous to the *RBL2* transcriptional cell cycle regulator in the IrxB region (see Fig. 12.4). This suggests that these two genes are likely to be remnants of an extra whole genome duplication thought to have occurred early in the teleost-lineage (Vandepoele *et al.*, 2004), after which, probably due to nonfunctionalization, the copy of *Q9Y2K8* on scaffold_28 was lost. These examples suggest that at least in these cases, and most probably for all non-*trans-dev* genes, the location of CNEs in their UTRs is merely incidental and plays no role in their regulation but rather regulates other genes in the vicinity. Like CNEs in general, there is a tight correlation of UTR CNEs with *trans-dev* genes, but there are no known post-transcriptional regulatory elements in UTRs that are as large or as well conserved as CNEs. Therefore, the location here is also likely to be incidental to their function, which in these cases appears to be *cis*-regulation.

XII. CNEs ARE LOCATED LARGE DISTANCES AWAY FROM THEIR PUTATIVE TARGET GENE

Distal CRMs are known to act at large distances from their target gene. Traditionally, prior to the availability of whole genomes, searches for such elements have occurred relatively close to the gene because of limitations in data as well as the concern of mis-assigning elements from further distances. The current distance limit between a target developmental gene and experimentally verified enhancers is ~1 Mb, for example, *SHH* (Lettice *et al.*, 2003), *SOX9* (Bishop *et al.*, 2000), and *MAF* (Jamieson *et al.*, 2002). CNEs represent excellent candidates for CRMs since many are located in the vicinity of a single *trans-dev* target gene, which provides an opportunity to gauge the distance limits over which such elements may act. Using only those elements conserved in *Fugu*, mouse, rat, and human, we examined the distribution of distances between CNEs and the gene TSS for all four species (Fig. 12.5). In the mammalian genomes, over a quarter of elements are within 100 kb of their target gene and over 88% are within 1 Mb. Incredibly, over 11% of elements were located between 1 Mb and 2 Mb from the target gene and ~1% are over 2 Mb, exceeding current limits for any known CRMs. A relative shift toward longer distance bins is seen in *Fugu* when compared to mammals, possibly reflecting the decreased level of compaction seen in a number of regions, although many regions are more compact than expected. This suggests that despite the tendency for compaction in the *Fugu* genome, it may be constrained by spatial requirements for elements interspersed among the CNEs identified here, including those that may be fish or species-specific CRMs. The furthest distance between a CNE and its likely target gene is 2.54–2.58 Mb in the mammalian genomes and 358.7 kb in *Fugu*. This is a 98 bp element (CRCNE00002604) located 3′ of the transcription factor *BCL11A*. The *bcl11a* gene is duplicated in *Fugu*, but the CNE is conserved only in one copy downstream of *bcl11a.1* (data from CONDOR). *BCL11A* is characterized by a large gene desert downstream containing over 100 CNEs with 13 CNEs located further than 2 Mb away from the gene. Regions containing *BCL11B*, *SOX11*, *TFAP2A*, and *NR2F1* also have CNEs located in excess of 2 Mb from these genes in mammals, although the distances vary in *Fugu* from 113.8 kb to 278.2 kb, reflecting asymmetrical rates of compaction previously noted. In addition, 42% (28/67) of single *trans-dev* genes have CNEs more than 1 Mb away in mammals. These data reflect those of Vavouri *et al.* (2006), who carried out a similar analysis but focused on duplicated CNEs that can be uniquely associated with paralogous genes in the human genome. One of the consequences

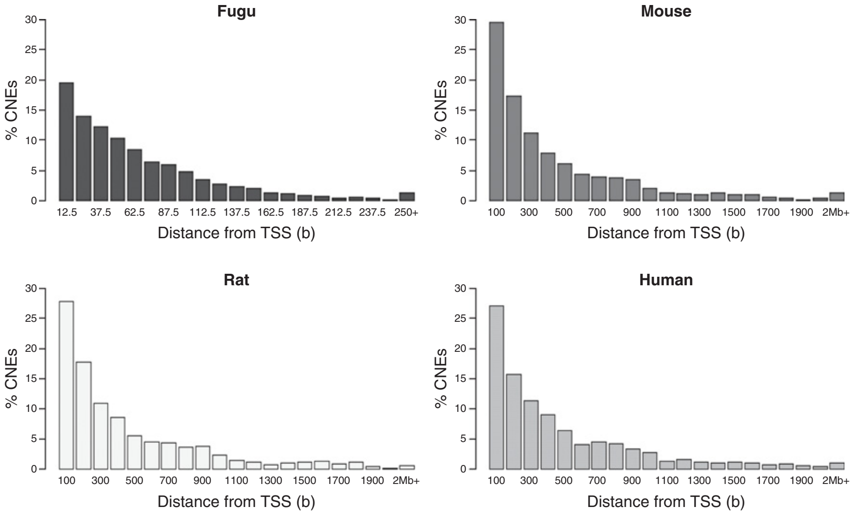

Figure 12.5. Proportion of CNEs at distances from the CNE to the transcription start site (TSS) of their target *trans-dev* gene in *Fugu*, mouse, rat, and human. Distance in mammals is measured at increments of 100 kb (i.e., 100 = 0–100, 200 = >100–200, etc.). Distance bins in *Fugu* are measured in increments of 12.5 kb (i.e., reflecting an average compaction rate of one-eighth in *Fugu* compared to mammalian genomes). The last bin in each graph contains all elements over 2 Mb away in mammals and 250 kb in *Fugu*.

of the very large distances across which CNEs act is that they often encompass neighboring genes. Invariably, this results in conservation not only of synteny but also of gene order in these regions (Ahituv *et al.*, 2005; Goode *et al.*, 2005; Kikuta *et al.*, 2007; Spitz *et al.*, 2003).

A. Mutation events in and around CNEs over evolution

Mutation events such as insertions/deletions (*indels*), inversions, and translocations are a major feature of genome evolution. Insertions and deletions involve changes in the sequence of the DNA, whil inversions and translocations act on a larger scale and involve predominantly structural rather than sequence changes. Given the distances over which CNEs are dispersed in the genome, it is likely, unless there is strong evolutionary selection against it, for CNEs to be involved in, or undergo, one or more of these events. Two approaches, BLAST and SLAGAN, have been used to identify CNEs that have undergone two forms of rearrangement, inversion and transposition.

1. Inversions

Mutational events involving inversions occur when a subsegment of DNA is removed from the sequence and then inserted back in the same location but in opposite orientation, often through repeat-mediated homologous recombination (Shaw and Lupski, 2004). CRMs, such as enhancers, have been shown to act both irrespective of orientation (e.g., Hill-Kapturczak *et al.*, 2003) and in an orientation-dependent manner (e.g., Swamynathan and Piatigorsky, 2002). It is therefore of interest to ascertain the proportion of elements which have undergone changes in orientation and whether this is a common feature of putative *cis*-regulatory elements. CNE orientation can be assessed relative to the transcriptional direction of the closest *trans-dev* gene in different species to identify inversions. We looked at orthologous CNEs in nine vertebrate genomes. Two hundred and twenty-four CNEs (~3.5% of this CNE set) were found to have changed orientation in one or more of these genomes, and are located around 50 of the 90 *trans-dev* regions in CONDOR. These changes are due to 114 individual inversion events, 40 of which involve two or more CNEs (Woolfe, 2006). Using orientation information from all genomes, it was possible to date these inversion events. A summary of this can be seen in Fig. 12.6.

The majority of the elements (62%) that have undergone changes in orientation are organism-specific (i.e., they have undergone inversion in only one genome), with the largest number occurring in frog and rat. In contrast to frog for which most of the changes in orientation are due to individual inversion events, the 43 CNEs showing changes in orientation in rat are the result of just three large-scale inversions. Two of these (containing all but one of the elements) are located in regions in the vicinity of *Barhl2* (Chr14) and *Sall3* (Chr18) genes. These regions are characterized by large gaps in the sequence that surround the inversion, suggesting that they are possibly due to errors in sequence assembly rather than a true inversion. Some 33% of inversion events appear to predate the mammalian lineage. Because of the incomplete nature of many vertebrate genomes, it is not always possible to identify the CNE and orthologous reference gene, or to locate both these features on the same scaffold, in order to confirm orthology. This was specially true of the genome of *Xenopus tropicalis*, which, like the *Fugu* genome, has no chromosomal mapping data and is highly fragmentary. Therefore, in many cases, dating of the inversion event could only be derived as the earliest point given the data available. Of elements with data available in all species, 10 derived from an inversion that occurred prior to the mammalian lineage but after the fish-tetrapod split. CNEs for which the change in orientation was seen between fish and tetrapods were placed at the root of the tree as in most cases it is not possible to know whether the inversion occurred within the fish lineage or early in the tetrapod lineage prior to the split with amphibians. A small number of these CNEs are duplicated in the *Fugu* genome, both from ancient vertebrate-specific

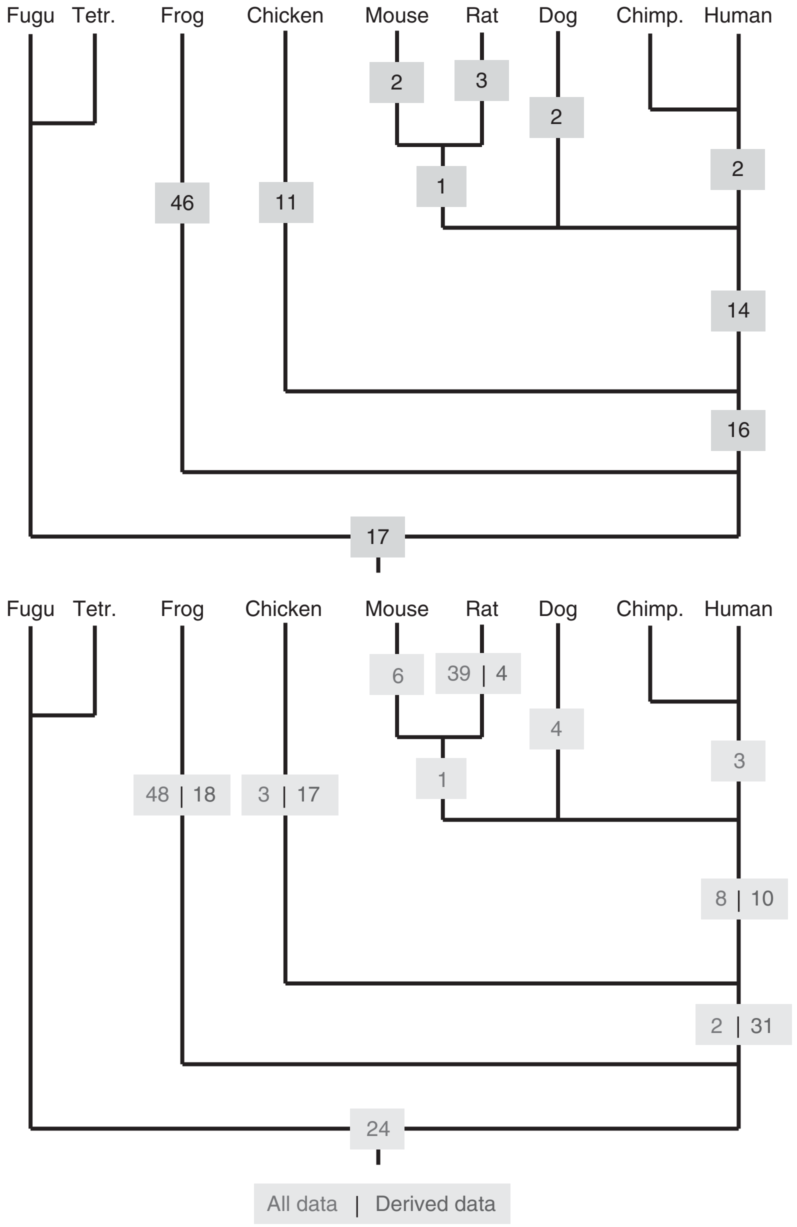

Figure 12.6. Inversion events involving CNEs in vertebrate genomes. Top tree—numbers in boxes represent individual inversion events involving one or more CNEs occurring at the evolutionary time point represented in the phylogenetic tree. Bottom tree—numbers

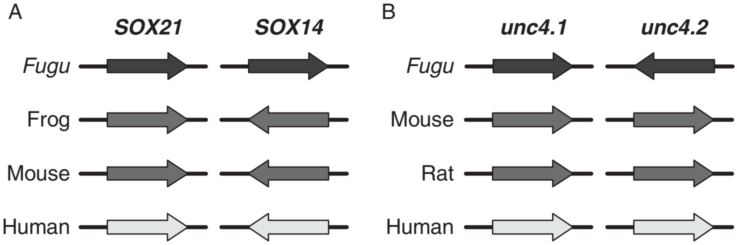

Figure 12.7. Changes in orientation between duplicated elements suggest both tetrapod-specific and *Fugu*-specific inversions. CNEs and their orientation are represented by block arrows. The relative orientation of the CNE is measured in relation to a reference *trans-dev* gene which is indicated above each set of arrows. (A) Two paralogous CNEs located upstream of *SOX21* and *SOX14* (CRCNE00000571 and CRCNE00000750) derive from whole genome duplication events at the origin of vertebrates (see Chapter 4, this volume). While the sequence of the CNE upstream of *SOX21* is found in the same relative orientation in all genomes, the same paralogous sequence in *SOX14* is orientated in opposite orientation in tetrapods whil remaining in the same orientation in fish, suggesting the inversion occurred early in tetrapod evolution. (B) *unc4.1* is a single copy gene in tetrapods but is duplicated in the *Fugu* genome and each gene region has a CNE [CRCNE00009841 (*unc4.1.2*) and CRCNE00009909 (*unc4.1.2)*] conserved to the same single-copy sequence in mammals. The CNE in *unc4.1.2* has undergone an inversion, indicating this is a fish- or *Fugu*-specific inversion event.

duplication events or duplications occurring early in teleost evolution providing a rare opportunity, through comparison, to date these inversions to the fish or tetrapod lineage. Two of these are highlighted in more detail in Fig. 12.7 and provide evidence that both fish-specific and tetrapod-specific inversion events have occurred around these elements and that the change in orientation has been conserved across evolution. This is strong evidence, at least for these elements, that CNEs may act as orientation-independent *cis*-regulatory elements.

2. Transpositions

In addition to inversions, transpositions are known to occur within genomes where a subsegment of DNA is removed from the sequence and then inserted back elsewhere in the genome in the same orientation and is thought to occur via mobile-repeat elements such as transposons. Changes in the order of a CNE

in boxes represent individual CNEs involved in inversion events. Where two numbers are side by side, the left hand figure represents the numbers where data is available at all time points in the tree. The figure on the right represents the best approximation to the inversion event in the absence of complete data.

between the *Fugu* and mouse, rat, dog (when utilized in the alignment), and human genomes were identified for all CNEs in CONDOR (Woolfe, 2006). Of the 6232 CNEs (224 elements that had undergone inversion were excluded from this analysis) representing 18,028 individual elements in these four mammalian genomes, just 7 CNEs were found to have undergone transposition events. Three derived from a single transposition event specific to rat within the gene desert upstream of the gene *ZFHX4*. Two are specific to mouse and occur upstream of *Fidgetin* gene and *Znf503*. One element appears to have undergone two separate transposition events in the mouse and rat genomes, which given the extremely low rate of transposition in other CNEs is highly unusual. There are no transpositions common to rodents or mammals, suggesting these events occurred after these organisms diverged. The transposed CNEs have a range of sizes between 45 bp and 180 bp and are significantly conserved in single copy in orthologous regions throughout the tetrapod lineage, indicating that they are not sequences identified by chance. Assuming the genomes in these regions have been assembled correctly, the extremely low rate of transposition involving CNEs is possibly due to deleterious effects of moving CNEs away from their target gene(s). It could also indicate that the order of elements along the sequence is of functional importance. If the isolated cases of transposition described above involve CNEs that represent true functional *cis*-regulatory elements, they are likely to have been fixed due to their fortuitous transposition to the same region in proximity to their target gene. Interestingly, more recent transpositions of conserved sequences have been identified, although these are not associated with *trans-dev* genes and are not conserved in fish genomes (Xie *et al.*, 2006; Bejerano *et al.*, 2006).

XIII. DISCUSSION

The identification and characterization of sequences responsible for the tight and precise regulation of vertebrate development remain one of the main challenges of the post-genomic era. In contrast to other functional elements such as coding exons, which have been the subject of intense study for many decades, our knowledge of the language, mode of action, and evolution of *cis*-regulatory elements are scant, in part because they are difficult to identify. The availability of ~6500 CNEs conserved between *Fugu* and mammals and located in the vicinity of key developmental regulator genes across a range of genomic environments provides the opportunity to study a large but focused noncoding cohort. These elements overlap no known coding transcripts or ncRNAs and are therefore likely to be highly enriched for putative *cis*-regulatory elements, providing the first opportunity to study their sequence and evolutionary character on a large scale.

The majority of CNEs are conserved across all currently available vertebrate genomes with the likelihood that missing CNEs are due to incompleteness of sequence coverage. Despite deep sequence conservation in the highly diverged genomes of fish and tetrapods, no similarity to invertebrate sequences can be found. A significant proportion of the *trans-dev* genes in this study has an ancient origin and is homologous to those identified in nonvertebrate chordates such as the sea squirt and in invertebrates such as flies and worms. Nevertheless, pairwise local alignments and more sensitive multiple alignments of orthologous noncoding DNA surrounding these genes do not identify any similarity outside the coding regions, even for the most highly conserved CNEs such as those found around *SOX21*. Thus it appears unlikely that the same set of sequences that appear to regulate vertebrate *trans-dev* genes are found in invertebrates, although intriguingly parallel sets of noncoding regulatory sequences have been identified around similar gene sets in nematodes (Vavouri *et al.*, 2007). A number of CNEs can be identified within a limited set of sequence read from the genome of the cartilaginous elephant fish, which, like other members of chondrichthyes such as sharks and rays, diverged prior to the split between Actinopterygii (ray-finned fishes) and Sarcoptergii (lobe-finned fishes) 480–570 million years ago (Venkatesh *et al.*, 2005). Thus far, it is possible to date the origin and fixation of these sequences in an ancestral genome sometime after invertebrate chordates but prior to the evolution of chondrichthyes 500–650 million years ago.

CNEs are extraordinarily constrained within the tetrapod lineage (with most CNEs ranging between 90% and 98% identity compared to the human sequence) despite being separated by up to 350 million years of evolution. In contrast, CNEs within teleost genomes appear to have evolved relatively rapidly both in comparison to their tetrapod orthologues and within teleost species. A faster rate of evolution in teleost genomes in comparison to mammalian genomes has been observed previously in relation to the neutral rate, which was found to be up to four times faster in the puffer fish than in mammals (Aparicio *et al.*, 2002, Jaillon *et al.*, 2004). This is highlighted most clearly by the significantly higher level of conservation observed in CNEs between elephant fish CNEs and human and teleost CNEs, despite the greater phylogenetic divergence (Venkatesh *et al.*, 2006). Teleosts, unlike sharks, have undergone an additional whole genome duplication event (Taylor *et al.*, 2003), which may be a reason for their elevated rate of evolution.

The CNEs discussed in this chapter cover a wide range of sizes (30–869 bp) and conservation (58–97%) between the *Fugu* and mammalian sequences. These attributes appear to be directly associated with their function as size and conservation of CNEs is not influenced by their genomic environment. For most length categories, CNEs have lower or equal levels of conservation than surrounding exons. This may not necessarily be true when compared against exons across the whole genome, as many of the exons within CNE regions derive from transcription factors and/or developmental genes which are known to be more highly conserved than other types of genes (Chuang and Li, 2004). CNEs of longer length (311–800 bp) generally exhibit higher sequence identity than exons of the same size, many CNEs have long stretches of complete identity between *Fugu* and mammalian sequences. Whil sequence conservation implies function, the precise function of the vast majority of CNEs remains unknown. A comparison of CNEs with protein coding exons through several analyses uncovers a number of revealing characteristics that shed light on their possible function. In the case of coding exons, the mechanism for conservation is known to be due to the constraint of the genetic code in specifying sequence coding for proteins. That most CNEs have little or no evidence of transcription suggests they do not form parts of mature transcripts, and this is consistent with a *cis*-regulatory role. A small proportion do match a large numbers of ESTs, similar to that seen in coding exons, indicating they may be missed, unannotated exons. However, some of these have proven roles as enhancers [e.g., CNEs in the Iroquois clusters (De la Calle-Mustienes *et al.*, 2005)], indicating that either the ESTs they match derive from genomic contamination or illegitimate transcription, or that some CNEs may act through transcribed intermediaries. Indeed, experiments on putative enhancers, located adjacent to the cytokine IL10 (Jones and Flavell, 2005) and the dlx-5/6 region (Feng *et al.*, 2006), demonstrate that a number of them transcribe intergenic RNA, leading to speculation that these regulatory elements function via an intermediate regulatory RNA. There is also intriguing evidence from work on the α-globin cluster that intergenic transcription from specific elements is used to modify chromatin structure leading to gene activation (Gribnau *et al.*, 2000). There is also recent evidence that some conserved elements have been distributed in the genome by ancient retrotransposons and can play roles as both enhancers and alternatively spliced exons depending on their genomic context (Bejerano *et al.*, 2006). Finally, there is also a possibility that some CNEs may function as novel ncRNAs as ~10% overlap predictions of significant RNA secondary structure. Again a number of these have proven *cis*-regulatory roles (Woolfe *et al.*, 2005), and it is not clear what proportion of these predictions represents real ncRNAs. It is therefore possible that CNEs may function through a number of different mechanisms.

CNEs have a number of evolutionary characteristics which are also distinct from those of coding exons. Regions of highly constrained blocks in multiple alignments are often signals of functional requirements. Coding exons show clear periodicity in the size of conserved block sizes indicative of the genetic code and the specificity of the first two bases of a triplet codon and the degeneracy of the third base. In contrast, CNEs show no periodicity in block sizes. However, CNEs do show a significant enrichment for blocks of size 7–13 bp in length, which correspond to the size of one to two transcription factor binding sites, commonly 5–10 bp in length (Woolfe, 2006). Intriguingly, CNEs also have a significant proportion of conserved block sizes of extreme length (>25 bp), which are, unsurprisingly, not as prevalent within block sizes for coding exons. This extreme conservation is analogous to that seen in "ultraconserved elements" defined as at least 200 bp of complete identity between human, mouse, and rat (Bejerano *et al.*, 2004), and indeed many of these sequences overlap CNEs, although this level of conservation is all the more surprising given the vast evolutionary distance involved. This degree of conservation cannot be fully explained by current models of CRMs, even for large numbers of transcription factor binding sites, as these binding sites are generally rather short and normally exhibit a level of redundancy. A number of ideas on the evolutionary mechanisms responsible for "ultraconservation" have been proposed, including increased DNA repair, multiple overlapping transcription factor binding sites, and decreased mutation (Bejerano *et al.*, 2004; Boffelli *et al.*, 2004). The last possibility is unlikely given the relatively rapid rate of evolution seen for many CNEs within teleosts, as well as a recent study that used single nucleotide polymorphism (SNP) data from the HapMap project and showed CNSs in the human genome are not located in mutation cold-spots but rather are selectively constrained (Drake *et al.*, 2006). It is also possible that if CNEs do bind transcription factors, these factors may have larger and more constrained binding site specificities than those currently known. It is clear therefore that until more is known about the mode of action of CNEs, the reasons behind these extreme patterns of conservation will remain both enigmatic and highly speculative.

Patterns of sequence insertion/deletion (indels) between *Fugu* and mammals are also distinct in CNEs compared to those in exons. Only a third of exons have any indels at all and those that do occur show a clear third base periodicity. This is, of course, due to the constraints of the genetic code which does not allow insertions of sequence lengths which would cause frameshift mutations. In contrast, almost 80% of CNEs have at least one indel, but these are mostly small and they follow a negative exponential distribution (i.e., larger indels are increasingly rare) (Woolfe, 2006). Nevertheless, many of the indels occur within some of the most highly constrained sections of the CNEs where we might predict changes would be most likely to have critical effects on function. It therefore appears that CNEs may, in some cases, be more accommodating to sequence indels than their highly constrained nature suggests. This was recently demonstrated in a functional study of an ultraconserved enhancer located within the introns of the *DACH1* gene (corresponding to CNE *CRCNE00000232* in CONDOR) (Poulin *et al.*, 2005). The central core of the enhancer has a region of 144 bp with 100% identity between *Fugu* and human and is critical to the function of the enhancer. However, the insertions of two nonfunctional 16 bp linkers into this core region caused no detectable modification of its *in vivo* activity, at least as far as the resolution of their transgenic mouse assay.

If we assume that CNEs represent modules for a number of protein binding sites, it is possible that sequence insertions can occur between binding sites without disrupting function and/or that some large CNEs represent a number of closely positioned but independent CRMs and that many of the large species-specific insertions lie at their boundaries. A whole range of further mutagenesis studies are required to see exactly which parts of a CNE can and cannot tolerate sequence insertions without disrupting their action and whether there is a limit to the distance at which parts of a CNE may be separated without affecting the function. This would indicate whether parts of a CNE are functionally dependent in some way or are completely independent.

CNEs are located within intergenic regions, the introns of *trans-dev* genes as well as the introns of neighboring non-*trans-dev* genes. There is no evidence that CNEs have any consistent bias in their position around their likely target gene consistent with the fact that many enhancers can act in a position-independent way. CNEs also appear to act, within limits, in a distance-independent manner. This is revealed through the simple fact that CNEs located many hundreds of kilobases away from their likely target gene in mammals are located at much shorter distances from the orthologous gene in the *Fugu* genome, and yet we assume that they function in the same way. This is consistent with current models of enhancer action which see the enhancer interact directly with the transcriptional machinery by the looping out of intervening DNA. Over 11% of human CNEs are located at distances from their likely target gene that exceed the distance limits over which currently experimentally verified enhancers are known to act (~1 Mb). This has major implications in the search for regulatory elements implicated in human disease. An increasing number of conditions are now being associated with SNPs or chromosomal break points/deletions that occur within noncoding regions at significant distances from a disease-linked gene (see other chapters in this volume, Kleinjan and van Heyningen, 2005). The knowledge that many critical regulatory elements may lie beyond these genomic disruptions will help in identifying disease-candidates. For example, Townes-Brock syndrome (TBS) is an autosomal dominant developmental disorder characterized by anal and thumb malformations and by loss of hearing. While most mutations for the disease map to the coding sequence of the transcription factor *SALL1*, a translocation breakpoint 180 kb upstream of *SALL1* on human chromosome 16 has also been linked to the TBS (Marlin *et al.*, 1999). Using CONDOR, 43 out of a total of 71 CNEs identified around *SALL1* map beyond this breakpoint, suggesting the removal of many critical regulatory elements away from the *SALL1* transcriptional unit is the cause of TBS in these cases.

The likelihood that CNEs work in a distance-independent manner does not preclude the possibility that the order of CNEs across the genome has important functional implications. Over 96% of all the CNEs in this study are conserved not only in sequence but have remained identical in order along the

DNA strand across vertebrate evolution. This suggests that some CNEs might play indirect regulatory roles that are position-dependent, such as structuring the genomic architecture around *trans-dev* genes through control of chromatin conformation or by acting as insulators that shield surrounding genes from the effects of enhancer action. The extremely low rate of CNE transposition suggests evolutionary constraint against the insertion and movement of transposons in these regions. A study of rearrangement events in the human and mouse genomes since their divergence showed that inversions and transpositions are a common occurrence (Kent *et al.*, 2003). There are a number of regions in the genome, though, which are lacking in any transposons [known as transposon free regions (TFRs)] which are enriched for gene regions containing genes with *trans-dev* ontologies (Simons *et al.*, 2006). It is therefore likely that constraint against the insertion and proliferation of transposons derives from the presence of large numbers of critical *cis*-regulatory elements which could become disrupted if transposed away from their target gene(s). Localized inversion events are also relatively rare within this set of CNEs, although they are more common than transpositions, possibly due to the orientation-independent nature of some enhancer elements.

In conclusion, the identification of thousands of potentially critical regulatory elements in the human genome has created an invaluable data resource. The organization of these elements throughout the genome has major implications in how future studies will now approach gene regulation and the characterization of regulatory disorders. Furthermore, knowledge of the function and distribution of CNEs around key developmental regulators will lead to a better understanding of developmental processes in vertebrates and lead to new paradigms of genetic disease.

References

Ahituv, N., Prabhakar, S., Poulin, F., Rubin, E. M., and Couronne, O. (2005). Mapping *cis*-regulatory domains in the human genome using multi-species conservation of synteny. *Hum. Mol. Genet.* **14,** 3057–3063.

Aloni, R., and Lancet, D. (2005). Conservation anchors in the vertebrate genome. *Genome Biol.* **6**(7), 115.

Aparicio, S., Morrison, A., Gould, A., Gilthorpe, J., Chaudhuri, C., Rigby, P., Krumlauf, R., and Brenner, S. (1995). Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, Fugu rubripes. *Proc. Natl. Acad. Sci. USA* **92,** 1684–1688.

Aparicio, S., Hawker, K., Cottage, A., Mikawa, Y., Zuo, L., Venkatesh, B., Chen, E., Krumlauf, R., and Brenner, S. (1997). Organization of the Fugu rubripes Hox clusters: Evidence for continuing evolution of vertebrate Hox complexes. *Nat. Genet.* **16,** 79–83.

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

Bagheri-Fam, S., Ferraz, C., Demaille, J., Scherer, G., and Pfeifer, D. (2001). Comparative genomics of the SOX9 region in human and *Fugu rubripes*: Conservation of short regulatory sequence elements within large intergenic regions. *Genomics* **78,** 73–82.

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.

Bejerano, G., Lowe, C. B., Ahituv, N., King, B., Siepel, A., Salama, S. R., Rubin, E. M., Kent, W. J., and Haussler, D. (2006). A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* **441**(7089), 87–90.

Bishop, C. E., Whitworth, D. J., Qin, Y., Agoulnik, A. I., Agoulnik, I. U., Harrison, W. R., Behringer, R. R., and Overbeek, P. A. (2000). A transgenic insertion upstream of sox9 is associated with dominant XX sex reversal in the mouse. *Nat. Genet.* **26,** 490–494.

Boffelli, D., Nobrega, M. A., and Rubin, E. M. (2004). Comparative genomics at the vertebrate extremes. *Nat. Rev. Genet.* **5,** 456–465.

Brenner, S., Elgar, G., Sandford, R., Macrae, A., Venkatesh, B., and Aparicio, S. (1993). Characterization of the pufferfish (Fugu) genome as a compact model vertebrate genome. *Nature* **366,** 265–268.

Brudno, M., Chapman, M., Gottgens, B., Batzoglou, S., and Morgenstern, B. (2003a). Fast and sensitive multiple alignment of large genomic sequences. *BMC Bioinformatics* **4,** 66–77.

Brudno, M., Do, C. B., Cooper, G. M., Kim, M. F., Davydov, E., Green, E. D., Sidow, A., and Batzoglou, S. (2003b). LAGAN and Multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* **13,** 721–731.

Brudno, M., Malde, S., Poliakov, A., Do, C. B., Couronne, O., Dubchak, I., and Batzoglou, S. (2003c). Glocal alignment: Finding rearrangements during alignment. *Bioinformatics* **19,** i54–i62.

Chuang, J. H., and Li, H. (2004). Functional bias and spatial organization of genes in mutational hot and cold regions in the human genome. *PLoS Biol.* **2,** E29.

Couronne, O., Poliakov, A., Bray, N., Ishkhanov, T., Ryaboy, D., Rubin, E., Pachter, L., and Dubchak, I. (2003). Free Full Text Strategies and tools for whole-genome alignments. *Genome Res.* **13**(1), 73–80.

De la Calle-Mustienes, E., Feijoo, C. G., Manzanares, M., Tena, J. J., Rodriguez-Seguel, E., Letizia, A., Allende, M. L., and Gomez-Skarmeta, J. L. (2005). A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts. *Genome Res.* **15,** 1061–1072.

Davidson, E. H., Rast, J. P., Oliveri, P., Ransick, A., and Calestani, C. (2002). A genomic gene regulatory network for development. *Science* **295,** 1669–1678.

Dermitzakis, E. T., Reymond, A., Scamuffa, N., Ucla, C., Kirkness, E., Rossier, C., and Antonarakis, S. E. (2003). Evolutionary discrimination of mammalian conserved nongenic sequences (CNGs). *Science* **302,** 1033–1035.

Dermitzakis, E. T., Reymond, A., and Antonarakis, S. E. (2005). Conserved non-genic sequences—an unexpected feature of mammalian genomes. *Nat. Rev. Genet.* **6,** 151–157.

Drake, J. A., Bird, C., Nemesh, J., Thomas, D. J., Newton-Cheh, C., Reymond, A., Excoffier, L., Attar, H., Antonarakis, S. E., and Dermitzakis, E. T. (2006). Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nat. Genet.* **38,** 223–227.

Dubchak, I., Brudno, M., Loots, G. G., Pachter, L., Mayor, C., Rubin, E. M., and Frazer, K. A. (2000). Active conservation of noncoding sequences revealed by three-way species comparisons. *Genome Res.* **10,** 1304–1306.

Duret, L., and Bucher, P. (1997). Searching for regulatory elements in human noncoding sequences. *Curr. Opin. Struct. Biol.* **7,** 399–406.

Durick, K., Mendlein, J., and Xanthopoulos, K. G. (1999). Hunting with traps: Genome-wide strategies for gene discovery and functional analysis. *Genome Res.* **9,** 1019–1025.

Elgar, G., Sandford, R., Aparicio, S., Macrae, A., Venkatesh, B., and Brenner, S. (1996). Small is beautiful: Comparative genomics with the pufferfish (Fugu rubripes). *Trends Genet.* **12,** 145–150.

Ellingsen, S., Laplante, M. A., Konig, M., Kikuta, H., Furmanek, T., Hoivik, E. A., and Becker, T. S. (2005). Large-scale enhancer detection in the zebrafish genome. *Development* **132,** 3799–3811.

Feng, J., Bi, C., Clark, B. S., Mady, R., Shah, P., and Kohtz, J. D. (2006). The Evf-2 noncoding RNA is transcribed from the Dlx-5/6 ultraconserved region and functions as a Dlx-2 transcriptional coactivator. *Genes Dev.* **20,** 1470–1484.

Frazer, K. A., Tao, H., Osoegawa, K., de Jong, P. J., Chen, X., Doherty, M. F., and Cox, D. R. (2004). Noncoding sequences conserved in a limited number of mammals in the SIM2 interval are frequently functional. *Genome Res.* **14,** 367–372.

Ghanem, N., Jarinova, O., Amores, A., Long, Q., Hatch, G., Park, B. K., Rubenstein, J. L., and Ekker, M. (2003). Regulatory roles of conserved intergenic domains in vertebrate Dlx bigene clusters. *Genome Res.* **13,** 533–543.

Goode, D. K., Snell, P., Smith, S. F., Cooke, J. E., and Elgar, G. (2005). Highly conserved regulatory elements around the SHH gene may contribute to the maintenance of conserved synteny across human chromosome 7q36.3. *Genomics* **86,** 172–181.

Göttgens, B., Barton, L. M., Gilbert, J. G., Bench, A. J., Sanchez, M. J., Bahn, S., Mistry, S., Grafham, D., McMurray, A., Vaudin, M., Amaya, E., and Bentley, D. R. (2000). Analysis of vertebrate SCL loci identifies conserved enhancers. *Nat. Biotechnol.* **18,** 181–186.

Gribnau, J., Diderich, K., Pruzina, S., Calzolari, R., and Fraser, P. (2000). Intergenic transcription and developmental remodeling of chromatin subdomains in the human beta-globin locus. *Mol Cell* **5,** 377–386.

Griffin, C., Kleinjan, D. A., Doe, B., and van Heyningen, V. (2002). New 3′ elements control Pax6 expression in the developing pretectum, neural retina and olfactory region. *Mech. Dev.* **112,** 89–100.

Hardison, R. C., Oeltjen, J., and Miller, W. (1997). Long human-mouse sequence alignments reveal novel regulatory elements: A reason to sequence the mouse genome. *Genome Res.* **7,** 959–966.

Hill-Kapturczak, N., Sikorski, E., Voakes, C., Garcia, J., Nick, H. S., and Agarwal, A. (2003). An internal enhancer regulates heme- and cadmium-mediated induction of human heme oxygenase-1. *Am. J. Physiol. Renal Physiol.* **285,** F515–F523.

Hoegg, S., Brinkmann, H., Taylor, J. S., and Meyer, A. (2004). Phylogenetic timing of the fish specific genome duplication correlates with the diversification of teleost fish. *J. Mol. Evol.* **59,** 190–203.

Jaillon, O., Aury, J. M., Brunet, F., Petit, J. L., Stange-Thomann, N., Mauceli, E., Bouneau, L., Fischer, C., Ozouf-Costaz, C., Bernot, A., Nicaud, S., and Jaffe, D. (2004). Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature* **431,** 946–957.

Jamieson, R. V., Perveen, R., Kerr, B., Carette, M., Yardley, J., Heon, E., Wirth, M. G., van Heyningen, V., Donnai, D., Munier, F., and Black, G. C. (2002). Domain disruption and mutation of the bZIP transcription factor, MAF, associated with cataract, ocular anterior segment dysgenesis and coloboma. *Hum. Mol. Genet.* **11,** 33–42.

Jones, E. A., and Flavell, R. A. (2005). Distal enhancer elements transcribe intergenic RNA in the IL-10 family gene cluster. *J. Immunol.* **175,** 7437–7446.

Kammandel, B., Chowdhury, K., Stoykova, A., Aparicio, S., Brenner, S., and Gruss, P. (1999). Distinct cis-essential modules direct the time-space pattern of the Pax6 gene activity. *Dev. Biol.* **205,** 79–97.

Kent, W. J., Baertsch, R., Hinrichs, A., Miller, W., and Haussler, D. (2003). Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci. USA* **100,** 11484–11489.

Kikuta, H., Laplante, M., Navratilova, P., Komisarczuk, A. Z., Engstrom, P. G., Fredman, D., Akalin, A., Caccamo, M., Sealy, I., Howe, K., Ghislain, J., Pezeron, G., *et al.* (2007). Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res.* **17,** 545–555.

Kleinjan, D. A., and van Heyningen, V. (2005). Long-range control of gene expression: Emerging mechanisms and disruption in disease. *Am. J. Hum. Genet.* **76,** 8–32.

Koide, T., Hayata, T., and Cho, K. W. (2005). Xenopus as a model system to study transcriptional regulatory networks. *Proc. Natl. Acad. Sci. USA* **102,** 4943–4948.

Lettice, L. A., Heaney, S. J., Purdie, L. A., Li, L., de Beer, P., Oostra, B. A., Goode, D., Elgar, G., Hill, R. E., and de Graaff, E. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* **12,** 1725–1735.

Levine, R., and Davidson, E. H. (2005). Gene regulatory networks for development. *Proc. Natl. Acad. Sci. USA* **102,** 4936–4942.

Loose, M., and Patient, R. (2004). A genetic regulatory network for Xenopus mesendoderm formation. *Dev. Biol.* **271,** 467–478.

Loots, G. G., Locksley, R. M., Blankespoor, C. M., Wang, Z. E., Miller, W., Rubin, E. M., and Frazer, K. A. (2000). Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288,** 136–140.

Marlin, S., Blanchard, S., Slim, R., Lacombe, D., Denoyelle, F., Alessandri, J. L., Calzolari, E., Drouin-Garraud, V., Ferraz, F. G., Fourmaintraux, A., Philip, N., and Toublanc, J. E. (1999). Townes-Brocks syndrome: Detection of a SALL1 mutation hot spot and evidence for a position effect in one patient. *Hum. Mutat.* **14,** 377–386.

Margulies, E. H., Blanchette, M., Haussler, D., and Green, E. D. (2003). Identification and characterization of multi-species conserved sequences. *Genome Res.* **13,** 2507–2518.

Margulies, E. H., Chen, C. W., and Green, E. D. (2006). Differences between pair-wise and multisequence alignment methods affect vertebrate genome comparisons. *Trends Genet.* **22,** 187–193.

Marshall, H., Studer, M., Popperl, H., Aparicio, S., Kuroiwa, A., Brenner, S., and Krumlauf, R. (1994). A conserved retinoic acid response element required for early expression of the homeobox gene Hoxb-1. *Nature* **370,** 567–571.

Martinez-Cruzado, J. C., Swimmer, C., Fenerjian, M. G., and Kafatos, F. C. (1988). Evolution of the autosomal chorion locus in *Drosophila.* I. General organization of the locus and sequence comparisons of genes s15 and s19 in evolutionary distant species. *Genetics* **119,** 663–677.

Mignone, F., Grillo, G., Liuni, S., and Pesole, G. (2003). Computational identification of protein-coding potential of conserved sequence tags through cross-species evolutionary analysis. *Nucleic Acids Res.* **31,** 4639–4645.

Miles, C., Elgar, G., Coles, E., Kleinjan, D. J., van Heyningen, V., and Hastie, N. (1998). Complete sequencing of the Fugu WAGR region from WT1 to PAX6: Dramatic compaction and conservation of synteny with human chromosome 11p13. *Proc. Natl. Acad. Sci. USA* **95,** 13068–13072.

Müller, F., Blader, P., and Strahle, U. (2002). Search for enhancers: Teleost models in comparative genomic and transgenic analysis of *cis* regulatory elements. *Bioessays* **24,** 564–572.

Nelson, J. S. (1994). "Fishes of the World" 3rd edn., John Wiley and Sons, New York, USA.

Nobrega, M. A., Ovcharenko, I., Afzal, V., and Rubin, E. M. (2003). Scanning human gene deserts for long-range enhancers. *Science* **302,** 413.

Nobrega, M. A., Zhu, Y., Plajzer-Frick, I., Afzal, V., and Rubin, E. M. (2004). Megabase deletions of gene deserts result in viable mice. *Nature* **431**(7011), 988–993.

Ochoa-Espinosa, A., Yucel, G., Kaplan, L., Pare, A., Pura, N., Oberstein, A., Papatsenko, D., and Small, S. (2005). The role of binding site cluster strength in Bicoid-dependent patterning in Drosophila. *Proc. Natl. Acad. Sci. USA* **102,** 4960–4965.

Oeltjen, J. C., Malley, T. M., Muzny, D. M., Miller, W., Gibbs, R. A., and Belmont, J. W. (1997). Large-scale comparative sequence analysis of the human and murine Bruton's tyrosine kinase loci reveals conserved regulatory domains. *Genome Res.* **7,** 315–329.

Ovcharenko, I., Nobrega, M. A., Loots, G. G., and Stubbs, L. (2004). ECR Browser: A tool for visualizing and accessing data from comparisons of multiple vertebrate genomes. *Nucleic Acids Res.* **32**(suppl 2): W280–W286.

Ovcharenko, I., Loots, G. G., Nobrega, M. A., Hardison, R. C., Miller, W., and Stubbs, L. (2005). Evolution and functional classification of vertebrate gene deserts. *Genome Res.* **15,** 137–145.

Pennacchio, L. A., and Rubin, E. M. (2001). Genomic strategies to identify mammalian regulatory sequences. *Nat. Rev. Genet.* **2,** 100–109.

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.

Rowitch, D. H., Echelard, Y., Danielian, P. S., Gellner, K., Brenner, S., and McMahon, A. P. (1998). Identification of an evolutionarily conserved 110 base-pair cis-acting regulatory sequence that governs Wnt-1 expression in the murine neural plate. *Development* **125,** 2735–2746.

Sandelin, A., Bailey, P., Bruce, S., Engstrom, P. G., Klos, J. M., Wasserman, W. W., Ericson, J., and Lenhard, B. (2004). Arrays of ultraconserved noncoding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics* **5,** 99.

Sanges, R., Kalmar, E., Claudiani, P., D'Amato, M., Muller, F., and Stupka, E. (2006). Shuffling of cis-regulatory elements is a pervasive feature of the vertebrate lineage. *Genome Biol.* **7,** R56.

Santagati, F., Gerber, J. K., Blusch, J. H., Kokubu, C., Peters, H., Adamski, J., Werner, T., Balling, R., and Imai, K. (2001). Comparative analysis of the genomic organization of Pax9 and its conserved physical association with Nkx2–9 in the human, mouse, and pufferfish genomes. *Mamm. Genome* **12,** 232–237.

Santagati, F., Abe, K., Schmidt, V., Schmitt-John, T., Suzuki, M., Yamamura, K., and Imai, K. (2003). Identification of cis-regulatory elements in the mouse Pax9/Nkx2–9 genomic region: Implication for evolutionary conserved synteny. *Genetics* **165,** 235–242.

Shaw, C. J., and Lupski, J. R. (2004). Implications of human genome architecture for rearrangement-based disorders: The genomic basis of disease. *Hum. Mol. Genet.* **13,** R57–R64.

Shi, W., Levine, M., and Davidson, B. (2005). Unraveling genomic regulatory networks in the simple chordate, Ciona intestinalis. *Genome Res.* **15,** 1668–1674.

Simons, C., Pheasant, M., Makunin, I. V., and Mattick, J. S. (2006). Transposon-free regions in mammalian genomes. *Genome Res.* **16,** 164–172.

Sironi, M., Menozzi, G., Comi, G. P., Cagliani, R., Bresolin, N., and Pozzoli, U. (2005). Analysis of intronic conserved elements indicates that functional complexity might represent a major source of negative selection on non-coding sequences. *Hum. Mol. Genet.* **14,** 2533–2546.

Sorek, R., and Safer, H. M. (2003). A novel algorithm for computational identification of contaminated EST libraries. *Nucleic Acids Res.* **31,** 1067–1074.

Spitz, F., Gonzalez, F., and Duboule, D. (2003). A global control region defines a chromosomal regulatory landscape containing the HoxD cluster. *Cell* **113,** 405–417.

Sun, H., Skogerbo, G., and Chen, R. (2006). Conserved distances between vertebrate highly conserved elements. *Hum. Mol. Genet.* **15,** 2911–2922.

Swamynathan, S. K., and Piatigorsky, J. (2002). Orientation-dependent influence of an intergenic enhancer on the promoter activity of the divergently transcribed mouse Shsp/alpha B-crystallin and Mkbp/HspB2 genes. *J. Biol. Chem.* **277,** 49700–49706.

Tautz, D. (2000). Evolution of transcriptional regulation. *Curr. Opin. Genet. Dev.* **10,** 575–579.

Taylor, J. S., Braasch, I., Frickey, T., Meyer, A., and van de Peer, Y. (2003). Genome duplication, a trait shared by 22,000 species of ray-finned fish. *Genome Res.* **13,** 382–390.

Thomas, J. W., Touchman, J. W., Blakesley, R. W., Bouffard, G. G., Beckstrom-Sternberg, S. M., Margulies, E. H., Blanchette, M., Siepel, A. C., Thomas, P. J., McDowell, J. C., Maskeri, B., Hansen, N. F., *et al.* (2003). Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424**(6950), 788–793.

Vandepoele, K., de Vos, W., Taylor, J. S., Meyer, A., and van de Peer, Y. (2004). Major events in the genome evolution of vertebrates: Paranome age and size differ considerably between ray-finned fishes and land vertebrates. *Proc. Natl. Acad. Sci.* **101,** 1638–1643.

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.

Vavouri, T., Walter, K., Gilks, W. R., Lehner, B., and Elgar, G. (2007). Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans. *Genome Biol.* **8,** R15.

Venkatesh, B., Tay, A., Dandona, N., Patil, J. G., and Brenner, S. (2005). A compact cartilaginous fish model genome. *Curr. Biol.* **15,** R82–R83.

Venkatesh, B., Kirkness, E. F., Loh, Y. H., Halpern, A. L., Lee, A. P., Johnson, J., Dandona, N., Viswanathan, L. D., Tay, A., Venter, J. C., Strausberg, R. L., and Brenner, S. (2006). Ancient noncoding elements conserved in the human genome. *Science* **314,** 1892.

Wasserman, W. W., Palumbo, M., Thompson, W., Fickett, J. W., and Lawrence, C. E. (2000). Human: Mouse genome comparisons to locate regulatory sites. *Nat. Genet.* **26,** 225–228.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S. E., and Attwood, J. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* **420,** 520–562.

Woolfe, A. (2006). Computational detection and analysis of putative *cis*-regulatory elements in vertebrate genomes. PhD Thesis, University of Cambridge.

Woolfe, A., Goodson, M., Goode, K., Snell, P., McEwen, G. K., Vavouri, T., Smith, S. F., North, P., Callaway, H., Kelly, K., Walter, K., and Abnizova, I. (2005). Highly conserved noncoding sequences are associated with vertebrate development. *PLoS Biol.* **3,** e7.

Woolfe, A., Goode, D., Cooke, J., Callaway, H., Smith, S., Snell, P., McEwen, G., and Elgar, G. (2007). CONDOR: A database resource of developmentally associated conserved non-coding elements. *BMC Dev. Biol.* **7**, 100.

Xie, X., Kamal, M., and Lander, E. S. (2006). A family of conserved noncoding elements derived from an ancient transposable element. *Proc. Natl. Acad. Sci. USA* **103,** 11659–11664.