Developmental Biology 350 (2011) 239–254

Contents lists available at ScienceDirect

Developmental Biology

journal homepage: www.elsevier.com/developmentalbiology

Review

When needles look like hay: How to find tissue-specific enhancers in model organism genomes

Maximilian Haeussler *, Jean-Stéphane Joly

U1126 MSNC INRA Group, UPR3294 NED, Institut Fessard, CNRS, 91 198 Gif-sur-Yvette, France

ARTICLE INFO

Article history:
Received for publication 14 April 2010
Revised 11 November 2010
Accepted 22 November 2010
Available online 3 December 2010

Keywords:
Cis-regulatory element
Fish
Zebrafish
Medaka
Transgenesis
Cis-regulation
Transcriptional regulation
Non-coding elements
Genome analysis
Enhancers

ABSTRACT

A major prerequisite for the investigation of tissue-specific processes is the identification of *cis*-regulatory elements. No generally applicable technique is available to distinguish them from any other type of genomic non-coding sequence. Therefore, researchers often have to identify these elements by elaborate *in vivo* screens, testing individual regions until the right one is found.

Here, based on many examples from the literature, we summarize how functional enhancers have been isolated from other elements in the genome and how they have been characterized in transgenic animals. Covering computational and experimental studies, we provide an overview of the global properties of *cis*-regulatory elements, like their specific interactions with promoters and target gene distances. We describe conserved non-coding elements (CNEs) and their internal structure, nucleotide composition, binding site clustering and overlap, with a special focus on developmental enhancers. Conflicting data and unresolved questions on the nature of these elements are highlighted. Our comprehensive overview of the experimental shortcuts that have been found in the different model organism communities and the new field of high-throughput assays should help during the preparation phase of a screen for enhancers. The review is accompanied by a list of general guidelines for such a project.

Introduction

Activating tissue-specific *cis*-regulatory elements – called "enhancers" (Banerji et al., 1981) – trigger gene expression in a given cell type, at the right developmental time and in the necessary quantity. They are tools of fundamental importance in diverse domains of biology. Cloned upstream of a fluorescent reporter gene, for example, they allow sorting of dissociated cells and tracking cell fate during embryogenesis with laser-scanning microscopes. They permit the analysis of essential genes by limiting the effect of functional assays to targeted cell populations: ectopic or over-expression of genes, knock-down with RNA interference or dominant-negative proteins or activation of Cre/Lox constructs can be performed in a tissue-specific manner. Finally, sequences of *cis*-regulatory elements can give clues about the trans-activating factor, helping to identify tissue-specific selector genes (Hobert, 2008).

In a more general sense, *cis*-regulatory elements also represent one big gap in our understanding of genomes, especially the huge non-coding parts: How much of the DNA is "junk" and what functions does the rest fulfill? What types of different functions are there? Which regions are implicated in human diseases (Kleinjan and Coutinho, 2009)? Although there are various types of *cis*-regulatory elements – reviewed by e.g. Arnosti (2003) and Maston et al. (2006) – robust assays are only available for enhancers. This is why the large-scale screens focus almost exclusively on these. Systemic tests have been conducted on one single locus at a time (Ishihara et al., 2008; Uchikawa et al., 2003) or on regions sampled from the whole genome (Woolfe et al., 2005). At the time of writing, the biggest project has screened around 1300 elements in thousands of mouse embryos (Pennacchio et al., 2006).

However, in many cases, researchers are interested in an element with a specific expression pattern . Given the size of the genome, relatively few regions have been tested already. Researchers are therefore often obliged to dissect the *cis*-regulatory landscape of a gene themselves and have to select a strategy on how to proceed. In the following, we provide guidelines for an enhancer screen targeting a single locus in a standard model organism. We summarize how various experimental improvements can be integrated in order to simplify the *in vivo* testing with transgenic model organisms. We describe some algorithms that predict the expression pattern from DNA sequences and point out their limits in the context of an enhancer screen. Finally, we highlight several topics deserving further investigation and comment on the importance of systematic *cis*-regulatory data collection.

* Corresponding author. Michael Smith Building, Faculty of Life Sciences, University of Manchester, Manchester M15 5RP, UK.
E-mail address: maximilianh@gmail.com (M. Haeussler).

0012-1606/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ydbio.2010.11.026

Main types of *cis*-regulatory elements and experimental testing

The high price of in vivo testing

To validate active individual regulatory elements, small DNA fragments (up to around 10 kbp) are cloned into plasmids one by one and tested for their activity with a reporter gene. As a result, elements active in tissues with available cell cultures are the ones best described in the literature. *In vivo*, however, there exists no experimental technique to screen large nucleotide sequences efficiently for their *cis*-regulatory potential at kilo base pair resolution. Complete testing of all randomly sheared fragments within a genomic locus is only feasible in simple model organisms such as ascidians or sea urchins (Keys et al., 2005; Cameron et al., 2004).

Nevertheless, protocols for other animals have been streamlined during the last years: observation of F0 embryos in mice is often sufficient (Loots, 2008). In zebrafish and nematodes, cloning can be avoided altogether by injecting PCR products (Woolfe et al., 2005; Hobert, 2002), although with an increase in mosaicism. In zebrafish, the number of assays can be reduced by testing genomic DNA from *Takifugu rubripes*, which is four times more compact while assumed to harbor similar regulatory elements (Barton et al., 2001). These experiments are still expensive in vertebrates, ranging between several hundred dollars per tested element in flies and fish to several thousand in mice (Table 1). For well-conserved elements, a time- and money-saving strategy might generate transgenic mice only with elements that have been tested first in fish. In many cases, conserved sequences yielded comparable expression patterns in both animals (Aparicio et al., 1995; Navratilova et al., 2009; Suster et al., 2009; Kimura-Yoshida et al., 2004).

If a given region does not fit into a plasmid, larger vectors, notably cosmids and BACs (Long and Miano, 2007), are more difficult to handle but can contain up to 40 kbp and 300 kbp of genomic data. Thanks to optimized protocols and better selectable markers, they can now be efficiently modified within 1 week (Sharan et al., 2009; Tursun et al., 2009; Smith, 2008; Venken et al., 2009; Ejsmont et al., 2009). Modifications of BACs (and more expensive knock-out mice) are the only way to remove elements from their context and find out if they are really necessary for a given expression pattern. Protocols and reagents are freely available from the National Cancer Institute at Frederick (NCICRF). Instead of screening individual DNA fragments to find the *cis*-regulatory element of interest, a BAC-clone with the gene replaced by a fluorescent protein coding sequence should often be sufficient to mark a cell type for subsequent analyses (Bouchard et al., 2005). Mouse lines for 800 BACs with an inserted GFP can be ordered through the GENSAT consortium (Geschwind, 2004). If large vectors are not an option, then individual regions in a locus have to be selected for testing.

Enhancer–promoter interactions

Where are enhancer elements found around a gene? In genes, a largely described location is in the introns, mainly in the first one. A handful of tissue-specific enhancers have also been described in coding regions. Several were described in 5′ untranslated regions (e.g. in the first exons of Pax6 (Zheng et al., 2001), IGF-1 (McLellan et al., 2006) and TH (Arányi et al., 2005)). Some have been recently discovered in translated coding exons (Hoxa2 (Tümpel et al., 2008; Lampe et al., 2008), Adamts5 (Barthel and Liu, 2008)). In addition, genome analyses found widespread non-coding selective pressure on coding regions (Woltering and Duboule, 2009; Chen and Blanchette, 2007; Kural et al., 2009) and exonic remnants after genome duplications and duplicated genes are known where all but one exon have disappeared (Dong et al., 2009b). This suggests that there might be more enhancers in transcribed and translated regions than is currently acknowledged but most are still expected to reside within the flanking non-coding regions around a gene or in introns within it.

The closest functional sequence here, directly upstream at around 50–100 bp, is the core or basal promoter (Juven-Gershon and Kadonaga, 2010). It used to be and still is often considered an essential but non-specific element of regulation, merely guiding the polymerase (Smale, 2001; Frith et al., 2008). Such a flexible structure with less sequence constraints might explain why the most conserved elements are located further upstream (Blanchette et al., 2006).Core promoters seem to be interchangeable between genes, as various studies in vertebrates have found a similar ratio of active enhancers although they used different core promoters (see Table 1).

But with more experimental data, the difference between the core promoter, the general “gateway to transcription” (Juven-Gershon et al., 2008), and tissue-specific elements has become less clear (summarized by e.g. Smale, 2001; Ohler and Wassarman, 2010). When assaying *cis*-regulatory sequences in invertebrates, not every enhancer could activate any promoter: for *Drosophila*, enhancers of gsb, gsbn, ant, bx require a certain type of promoter (DPE- or TATA-containing) (Li and Noll, 1994; Ohtsuki et al., 1998; Butler and Kadonaga, 2001). A mutation of the *yellow* or *oaf* promoters can change the interactions with enhancers (Lee and Wu, 2006; Merli et al., 1996). In *C. elegans*, a neural motif is not active when combined with non-neural promoters (Wenick and Hobert, 2004). Mammalian genome analyses found in roughly one third of the cases a relationship between the direct upstream sequence and the cell type where a gene is expressed (Smith et al., 2007; Roider et al., 2009; Vandenbon and Nakai, 2010). In cell cultures, the expression response to p53 depends on the type of basal promoter (Morachis et al., 2010) and specific transcription factors like E2F bind to a large proportion of all core promoter regions (Xu et al., 2007). In an extreme case, a tissue-specific element in sea urchin showed two different expression patterns, depending on the basal promoter it was combined with (Kobayashi et al., 2007).

The dependence on the basal promoter can lead to problems in medium-scale enhancer screens that test elements genome-wide, from various loci. In such experiments, non-coding fragments are combined with one standardized promoter, typically pHsp or pBeta-globin. In *Drosophila*, the possible incompatibilities motivated the development of an artificial *Super Core Promoter*, a mix of several different sequences with the goal of increased enhancer compatibility and high expression levels (Juven-Gershon et al., 2006). For some mammalian cell lines, optimized sequences have been synthesized that perform better than the CMV minimal promoter (Schlabach et al., 2010). In zebrafish, (Gehrig et al., 2009) analyzed almost all

Table 1

Organism	Delivery	Time from experiment to observation	Price transgenesis, academic rate	Source
D. melanogaster	injection	1–2 weeks	$250	the best Gene.com
C. elegans	injection	1–2 days	$150–$250	C. elegans Core, NTHU, Taiwan
C. intestinalis	electroporation	1 day	No core	
Zebrafish	injection	1–2 days	$450	Amagen Platform, CNRS, France
Chicken	electroporation (not all cells)	1 day	No core	
Mouse	injection	7–13 days	$2200	Ohio State Univ., Mouse Core

combinations of 19 core promoters and 11 enhancers by generating 202 constructs and scoring images of around 18.000 embryos. Based on their data, they can rank core promoters into more general ones (krt4, hsp70) and those that act differently with most enhancers (ndr1 and eng2b). One can see that in a screen, the endogenous core promoter should be preferred, if possible.

This is one reason why the most common starting point in the search for activating tissue-specific *cis*-regulatory elements is the region of the core promoter and its upstream sequences. When referring to the start of transcription of a gene, researchers should be aware that depending on the quality of genome annotation of a model species and the number of incomplete ESTs, the beginning of a transcript displayed by genome browser does not always reflect the real start of transcription. Some years ago, full-length cDNA sequencing showed that a third of human transcripts were actually starting further upstream (Suzuki et al., 2002) and systematic 5′ RACE indicates that this number could be higher (Denoeud et al., 2007). The phenomenon should be more common for less-annotated genomes, like zebrafish, frog or chicken. It is therefore advisable to check all available evidence (all possible gene models, aligned ESTs and also DBTSS (Wakaguri et al., 2008) in the case of human and mouse) when determining the 5′ end of a gene. One might consider running 5′ RACE experiments in the case of less-studied genes or organisms.

Human/rodent sequence analyses indicate that the region within 2–2.5 kbp of the gene start is under selective constraint (Keightley and Gaffney, 2003; Keightley et al., 2005). This can serve as a rough guideline for delimiting the boundaries of proximal fragments. For genes expressed in adult tissues, the sequence composition of the upstream region seems to follow different rules (Vandenbon and Nakai, 2010; Roider et al., 2009). Loots (2008) and Nelson et al. (2004) have observed that housekeeping genes or those with expression in differentiated tissues seem to keep their regulatory regions relatively close, while these regions tend to be more distant from genes with a developmental expression profile.

In the standard *in vivo* assay, a fragment upstream of the start of transcription (<10 kb) is cloned into a plasmid and the DNA then injected (mouse, fish, sea urchin, flies, nematodes) or electroporated (ascidians, chicken) into eggs, the gonads or embryos. For invertebrates with small genomes like *C. elegans*, *Drosophila* and *C. intestinalis*, one plasmid can contain a large part of the upstream region and very often reproduces the gene expression pattern faithfully (Boulin et al., 2006). The same approach worked for some vertebrate genes (e.g. Wang et al., 2002; Park et al., 2000; Yoshikawa et al., 2007). These proximal sequences already contain the correct endogenous basal promoter and should be the first region to test in any screen. But in larger genomes and especially in loci with long intergenic regions, only a small part fits into one construct. Therefore, many promoter regions recapitulate only a part or none of the wild-type expression pattern of a gene and a laborious search for more distant elements might be required.

Long-range interactions and position effects

In the early 1980s, mutations of the β-globin locus in Thalassemia patients, validated in model organisms (reviewed by e.g. West and Fraser, 2005), suggested that chromatin loopings permit long-range *cis*-regulatory contacts. *Cis*-regulatory elements were shown to be necessary for chromatin modifications located 100 kbp away from the gene (Forrester et al., 1990). Later, two methods refined these studies: chromatin conformation capture assays (Dekker et al., 2002) and chromatin stained by tagged RNAs (TRAP). The TRAP protocol involves targeting horseradish peroxidase activity to the primary transcripts. This technique leads to the deposition of a biotin tag on chromatin proteins in the immediate vicinity of the transcribed gene. Confocal microscopy (Carter et al., 2002) showed that β-globin enhancers are in close proximity with the neighboring basal promoters during target gene expression. The necessary DNA loopings are induced and anchored by transcription factors like GATA1 (Vakoc et al., 2005) and preceded by chromatin modifications of the globin enhancers (Li et al., 2006). Contacts like these can even reach out to other chromosomes and regulate genes in *trans* (Chen et al., 2002; Simonis et al., 2006; Lomvardas et al., 2006; Ronshaugen and Levine, 2004). In addition, analysis of the Shh locus with large vectors and again chromosomal rearrangements in patients showed that enhancers can be located up to 1 MB away from their target gene in mammals in extreme cases (Lettice et al., 2003) (reviewed by Long and Miano, 2007). Long distances – in relation to the total genome size – have also been reported in invertebrates (Jack et al., 1991; Dorsett, 1993; Conradt and Horvitz, 1999; Smith, 2008). On a more general level, based on around 100 transcription factor/developmental genes with elements conserved in fish and mammals, (Woolfe and Elgar, 2008) estimated that 50% of their distal regulatory elements are located >300 kb away in human (fugu: >50 kb). One can see that the radius of a regulatory element search can be extensive in certain cases. As noted above, genes expressed in only few or adult tissues are expected to have less distant enhancers, though we are not aware of extensive studies estimating this bias.

Apart from testing individual conserved non-coding elements, discussed in detail in the next section, one simple way to reduce the search space is gene synteny comparisons. Various authors have argued that long-range regulation should limit possible chromosomal rearrangements and maintain some exceptionally long and well-conserved syntenic blocks. Based on experimental data, this scenario has been cited as an explanation for enhancers located within neighboring genes of mouse Pax9 and Nkx2 (Santagati et al., 2003), zebrafish Shh (Goode et al., 2005) Pax6, Fgf8 and Rx (Kikuta et al., 2007), and amphioxus Pax1/9 (Wang et al., 2007a,b). Enhancers regulating distant genes were proposed to trigger the extreme conservation of the Hoxa/Hoxd loci in fish and mammals (Lee et al., 2006). Genome analyses found a correspondence between synteny and the distribution of conserved non-coding elements in tetrapods (Ahituv et al., 2005), flies (Engström et al., 2007) and amphioxus (Hufton et al., 2009). Following this model, synteny breakage could be used to delimit the boundaries of enhancer action, provided that expression of the genes in the locus is conserved at long evolutionary range. The idea is that when a given region is not flanking the ortholog in a related species, the enhancer is less likely to be located in this region. The genome browsers of UCSC and Ensembl provide a DNA-based synteny view for this ("UCSC Net Tracks" and "Ensembl Multicontigview"). Metazome (www.metazome.net) shows only genes, which makes it easier to use but less sensitive; the recent tools Synorth (Dong et al., 2009a,b) and Genomicus (Muffato et al., 2010) combine a view of a genomic locus with phylogenetic relationships of the genes in it. Fig. 1 shows an example of the gene SALL1 based on the UCSC Genome Browser (Kent et al., 2002) where the collinearity of non-coding fragments in *X. tropicalis* suggests that most enhancers concentrate in a 1.5 MB segment around the gene.

Long-range control can lead to problems of reproducibility when testing *cis*-regulatory elements. In most transgenic techniques, sequences and reporter genes are randomly inserted into the genome. The "position effect" (Spradling and Rubin, 1983) describes expression pattern variations between different transgenic animals due to the influence of the genomic context around the construct. In *Drosophila*, the effect between different genomic insertion sites can be 100-fold and RNAi constructs lead to very diverse wing phenotypes depending on the insertion site (Markstein et al., 2008). A common way to cope with this variability is to report only the common pattern between several transgenic embryos. An alternative is the addition of flanking insulators around the reporter construct (Potts et al., 2000; Markstein et al., 2008). In *Drosophila*, PhiC31 insertion surrounded by gypsy insulators allows to reduce the position effect (Ni et al., 2009). In zebrafish, a vector containing a minimal promoter, flanking

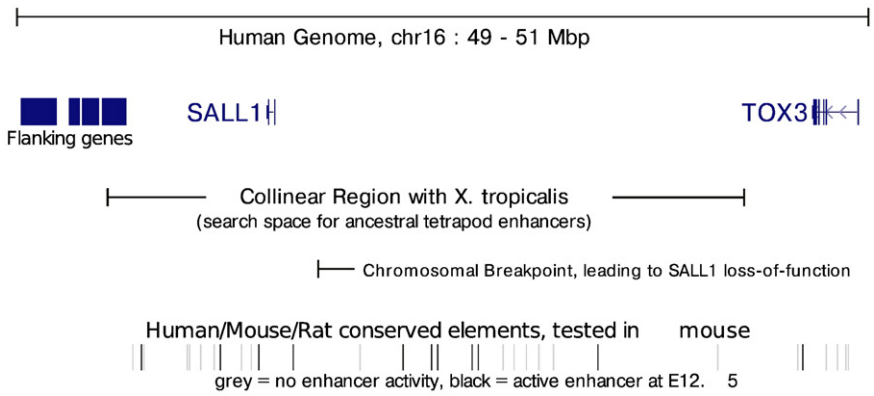

Fig. 1. Adapted from an annotated screenshot of the UCSC Genome Browser showing the SALL1-locus (hg18, chr16:49086985-51296445) with neighboring genes, inspired by Ahituv et al. (2005). The syntenic (collinear) region with X. tropicalis is limited to a 1.5 MB fragment around the gene SALL1. Since the expression pattern of Sall1 is conserved in tetrapods, an enhancer screen should concentrate on this 1.5 Mbp region and include the introns of a flanking gene. Two additional lines of evidence that support this hypothesis are shown: a) Chromosomal breakage of a major part of the non-coding region has the same effect as a mutation in the gene SALL1 (Marlin et al., 1999). b) Several conserved non-coding sequences in this region were tested in mice and direct an expression pattern reminiscent of SALL1 (Pennacchio et al., 2006).

insulators and a positive control is readily available (Bessa et al., 2009). In mice, constructs can be integrated into the transcriptionally "neutral" locus HPRT, now aided by a set of readily available plasmids (Yang et al., 2009). This is useful when one strives to quantify the effects of small changes in known *cis*-regulatory sequences (Ahituv et al., 2007b) but certainly too laborious in the context of a screen. When testing conserved elements, one should keep in mind that some *cis*-regulatory regions have been shown to activate several genes. These are called "global control regions" or "locus control regions," 20–30 kbp-long arrays of enhancers that regulate a cluster of diverse genes (Lower et al., 2009; Spitz et al., 2003). Well-known loci include, apart from the alpha- and beta-globins, the interleukins, and the Evx2-HoxD locus (reviewed by Spitz and Duboule, 2008). The influence of so-called global control regions/locus control regions on given regions of the genome results in "gene expression neighborhoods" (Oliver et al., 2002) or "regulatory landscapes" (Spitz and Duboule, 2008). Therefore, the experimenter has to be prepared to screen up to 1 MB of flanking sequence around the gene of interest, even beyond neighboring genes or within their introns, with coregulated genes nearby as potential additional targets and with an experimental technique that addresses potential position effects.

Cis-regulatory element interactions

Enhancer screens assume that each functional element can be tested individually. It is expected that for two enhancers combined in one plasmid, the resulting expression pattern is the sum of both individual patterns (see e.g. (Kirchhamer et al., 1996 and references therein; Visel et al., 2009a). But interactions between *cis*-regulatory elements are known: In the context of the genome, these interactions are modulated by insulators, which have been analyzed in detail for many years in the *Drosophila bithorax* complex and the *yellow* locus (reviewed by Akbari et al., 2006; Maeda and Karch, 2007). In vertebrates, insulators are assumed to be bound by CTCF (Kim et al., 2007), the main remodeller of chromatin loopings (Phillips and Corces, 2009).

Insulators are not the only modifiers: elements have been described with a repressor (Conte and Bovolenta, 2007) or amplificator effect on flanking enhancers (Yuh et al., 1998; Irvine et al., 2008; Kirchhamer et al., 1996 and references therein) or with both effects at the same time (Kulkarni and Arnosti, 2003; Nolis et al., 2009). As a result, a combination of enhancers is required to drive the correct expression pattern: Troponin muscle expression is modulated by secondary enhancers that are silent when tested individually (Guerrero et al., 2010). The endogenous expression pattern of the gene Shh could only be recreated with a certain set of elements, not any single one alone (Ertzer et al., 2007a,b). A model of the complete expression pattern of *even-skipped* in *Drosophila* requires 34 binding sites, taking into account repressors, competitive binding and quenching effects between different modules (Janssens et al., 2006). These examples illustrate the problem that it might be almost impossible to prove that an element is completely nonfunctional, as it would have to be tested in combination with potential interacting partners.

Non-coding elements: Conservation and activity

Non-coding conservation as a predictor of cis-regulatory function

The biggest help in finding short tissue-specific enhancers within loci of several mega base pairs is alignments with non-coding sequences from other species. Since their first analyses between human and mouse in the β-globin locus (Hardison and Miller, 1993) and later the mouse genome project (Hardison et al., 1997; Waterston et al., 2002) surprisingly many of these alignable sequences have been found. They are not an artifact of non-randomly distributed mutation rates in the genome (mutational cold-spots) but have been shown to be under selection (Drake et al., 2006; Casillas et al., 2007; Katzman et al., 2007; Sakuraba et al., 2008) although deep sequencing of one ultraconserved element indicated that it accumulated less mutations than flanking sequences in colorectal cancer samples (De Grassi et al., 2010). As a result, researchers have been concentrating on alignable conserved elements during the last years when searching for tissue-specific enhancers and this approach has been very successful. Today, the standard genome browsers (UCSC, Ensembl, Vista Genome Browser) allow the identification of CNEs with a mouse click. Depending on the filtering applied, these regions have been named differently: *conserved non-coding sequences* (CNS, >X% identity over Y bps) (Dubchak et al., 2000), *deeply conserved elements* (human/fish) (Attanasio et al., 2008), *ultraconserved* (200 bp identical human/mouse/rat) (Bejerano et al., 2004), *extremely conserved* (Pennacchio et al., 2006) or *extremely highly conserved sequences* (de la Calle-Mustienes et al., 2005), *hyperconserved sequences* (more than 5 nucleotides in five species (Guo et al., 2008)) and many more (reviewed by Woolfe and Elgar, 2008).

Table 2 shows a small subset of the numerous enhancer screens of non-coding conserved elements. To identify them, researchers used a combination of various species and rather simple cutoffs. Some general conclusions become apparent: the expression patterns of the elements varied a lot from one specimen to the other which can pose a challenge during imaging and interpretation. Therefore, bigger screens generally describe the expressing cell populations in much less detail. While one single medium-scale program (Pennacchio et al., 2006) has uncovered more enhancers than all other

Table 2
A selection of studies that describe tissue-specific elements identified by non-coding conservation. In the biggest screen in mouse embryos which were fixed at e12.5, only 497/1083 CNEs were active (45%). If this screen (low resolution) and chicken sequences (tissue-dependent electroporation) are not counted, out of 117 conserved non-coding sequences, 93 drove a tissue-specific expression pattern (80%).

Locus	Organism (DNA/org)	Sequence conserved with	Tissue or cell type	Tested enhancers	Confirmed enhancers	Position relative to gene	Trans-acting factor determined?	Promoter	Publication
Sall1	chicken	human	anterior neural ridge	5	1	intron	No	thymidine kinase	[Izumi et al., 2007]
Sox2	chicken	human	Di/mesencephalon, Nasal and otic placodes, Rhombencephalon, Neural induction, Head ectoderm Mesencephalon Spinal cord, Late lens, Dorsal root ganglia	25	10	50 kb 5′	No	Herpes virus thymidine kinase	[Uchikawa et al., 2003]
Eya1	chicken		Hensen's node, neural tube, migrating neural crest cells, otic vesicle, olfactory placode, cranial ganglia, trigeminal ganglia	29	10		many (match)	Herpes virus thymidine kinase	[Ishihara et al., 2008]
Dach1	mouse	fugu	fore/mid/hindbrain, retina, limb buds, neural tube, genital eminence	9	7	<870 kb	No	Hsp68	[Nóbrega et al., 2003]
Dlx1/2	mouse	zebrafish	anterior entopeduncular area, subventrivular zone, parvalbumin-, calretinin-, neuropeptide Y, and other interneurons	4	4	<12 kb	No	Hsp68	[Ghanem et al., 2007, 2003]
Flt4, PD FFrβ, Ece1, Nrp1, Foxp1	mouse	human?	endothelium	10	5	?	FoxC2, Ets (ectopic expr/KO)	β-globin	[De Val et al., 2008]
Gata2	mouse	human	rostral urogenital system, caudal urogenit. system	4	2	3′, 1 MB	No	Gata2	[Khandekar et al., 2004]
Hoxb4	mouse, fugu/mouse		rhombomer 7/8, anterior mesoderm, neural tube	3	3	intronic	No	Hsp68,Hoxb4	[Aparicio et al., 1995]
Hob2	mouse, chicken	bat, chicken	rhombomere 4	1	1	introns	HoxB1, Prx, Prep1 (emsa, mut, overexpr)	β-globin	[Maconochie et al., 1997]
Mbp	mouse		oligodendrocytes at different stages	4	4	15 kb 5′	Nkx (mut)	Hsp68	[Farhadi et al., 2003]
nicotinic acetylcholine receptors	mouse	human	adrenal gland, superior cervical ganglion, pineal gland, SCG neurons,	1	1	30 kb 5′	No	None (BAC deletion)	[Xu et al., 2006]
Nkx2-5	mouse	human	heart common atria, common ventricle, aortic sac, distal stomach region, tongue,	3	3	27 kb 5′	Gata/Smad (mut)	Hsp68	[Chi et al., 2005]
Otx2	fugu/mouse, fugu/zebrafish	mouse	roof of dienc., medio-caudal telenc., ventral dienc., ZLI, cephalic mesenchyme, trigeminal ganglions, cranial nerves, dorsal dienc., rhombenc., nasal pits, first branchial groove	7	7	60 kb	No	Otx2	[Kimura-Yoshida et al., 2004]
Pax6	mouse	human	late eye development, diencephalon (auto), heart, rhombencephalon	4	3	intronic	Pax6 (emsa)	Hsp68	[Kleinjan et al., 2004]
Pax6	human/zebrafish (same)	human	left and right habenulae, roofplate, pineal, medial habenulae	8	6	~300 kb 5′ and 3′	No	Gata2, Hsp70, Ngn1, Atpc11, Atpc11, Sox3	[Navratilova et al., 2009]
Shh	zebrafish/mouse	mouse	embryonic shield, hypothalamus, zli	3	3	introns	No	Gata2	[Ertzer et al., 2007a,b]
Sox10	mouse	chicken	otic vesicle, oligodendrocytes neural crest, peripheral nervous system, adrenal gland, sympathetic ganglia, neural crest	7	5	65 kb	Sites for Sox/lef/ Pax/Ap2 (EMSA)	Hsp70	[Werner et al., 2007]
Sox21, Pax6, Hlxb9, Shh	zebrafish	human	approx annotation: nervous sys., sens. organs, notochord, muscle, blood, heart, skin	25	23	various	No	β-globin	[Woolfe et al., 2005]
Sox3	human/zebrafish	zebrafish	brain, epiphysis, floor plate, inner ear, cerebellum	8	6	300 kb 3′, 100 kb 5′	No	Gata2 + 5 others	[Navratilova et al., 2009]
Various	zebrafish	human	Rough classification into 6 tissues, quantitative	16	10	various	No	cMLC2, luciferase	[Shin et al., 2005]
Various	mouse	human	Rough classification in fore/mid/hindbrain	1083	497	None		β-globin	[Pennacchio et al., 2006]

laboratories together, it is currently lacking a detailed annotation of the exact domains of embryonic expression; because a systematic histological analysis would take much time. Most multi-locus assays used basal promoters from Hsp68 or β-globin, underscoring their role as the standard basal promoters. The most common criteria for interspecies conservation are strong conservation in human/mouse, probably because of the quality and availability of their genomes. In total, the majority (80%) of CNEs in Table 2 showed a *cis*-regulatory effect.

Obviously, many conserved elements remain to be tested. Depending on the parameters, one can find between several hundred (ultraconserved), several thousands (human/fish) to several hundred thousands (mammals) of these in vertebrate genomes (Visel et al., 2007). Thanks to extensive sequence analyses some general properties have emerged (Box 1). These features have important implications when selecting candidate enhancer sequences: as regions are unevenly conserved depending on the locus, there is currently no "optimum" combination of genomes to find them. Therefore, the most conserved regions in a given locus should be given preference. Cross-species testing is limited to relatively closely related organisms, like human/mouse or human/fish.

Reasons for strong non-coding conservation

In the context of *cis*-regulatory non-coding elements, a puzzling question remains: how can an enhancer be conserved over 200 bp without a single base pair mutation between human and mouse (Bejerano et al., 2004), if a transcription factor binding site is only 4–8 base pairs long? Why do the nucleotides between the binding sites not mutate? Their only major characteristic seems that some of the most exceptional CNEs are derived from transposable elements (Nishihara et al., 2006; Xie et al., 2006; Bejerano et al., 2006), but this does obviously not account for their conservation.

One explanation could be a double function of the elements, if they serve as enhancers and regulatory RNA at the same time. Generally, many non-coding sequences are transcribed (Birney et al., 2007) and some of these transcripts could be functional, as their expression pattern corresponds to chromatin boundaries (Akbari et al., 2006; Rinn etal., 2007). Cross-species alignments suggest that thousands are structured RNAs (Washietl et al., 2007) and various examples have shown that non-coding RNAs can regulate directly the transcription process (reviewed by Amaral and Mattick, 2008). It is therefore not surprising that 70% of extragenic transcription start sites show an enhancer-like

Box 1

Main features of CNEs

- **Compared to exons/introns:** some CNEs (called ultraconserved sequences) are much better conserved than most protein coding sequences (Bejerano et al., 2004; Dermitzakis et al., 2003). The percentage of non-coding relative to coding sequences increases with organism complexity from yeast, worms and insects to vertebrates (Siepel et al., 2005). In flies, CNE sequences of intronic elements are indistinguishable from intergenic ones (Bergman and Kreitman, 2001).
- **Evolutionary distance and conservation:** CNEs have a "short lifetime" and are mostly phylum-specific: only 56 of the vertebrate sequences can be found in a cephalochordate, the amphioxus (Putnam et al., 2008). No single non-coding sequence is conserved between vertebrates on the one hand and flies, worms or ascidians, on the other (Bejerano et al., 2006), most CNEs are aligned separately within vertebrates, flies, ascidians and plants. Most mammalian CNEs seem to have emerged during the early tetrapod history (Stephen et al., 2008) and have been strongly retained during mammalian evolution (McLean and Bejerano, 2008). The best-conserved primate regions are also best conserved in mammalian genomes (Prabhakar et al., 2006; Wang et al., 2007a,b). The probability that an element shows *cis*-regulatory activity increases with its conservation (Pennacchio et al., 2006) and also with the density of surrounding elements (Prabhakar et al., 2006).
- **Composition:** CNEs show a biased A/T content with 6% more A/T than in the flanking regions, in vertebrates, worms and plants (Walter et al., 2005; Vavouri et al., 2007; Li et al., 2009). These A/Ts are clumped into stretches, while CpG contexts are avoided. They contain many sequences that look like transcription factor binding sites. However, of all over-represented motifs in them, only very few can be assigned to a characterized transcription factor family (Minovitsky et al., 2007).
- **Insertions/Deletions:** in five regions conserved in sea urchins, insertions >20 bp are almost absent (Cameron et al., 2005), but a 16 bp-insertion into one of the best-conserved enhancers in the genome, the Dach locus, did not change its expression pattern (Poulin et al., 2005). In flies, (Sinha and Siggia, 2005) found an excess of tandem repeat insertions relative to deletions in conserved regulatory regions. CNEs in flies are enriched in long ungapped blocks (>20 bps) (Papatsenko et al., 2006).
- **Duplications:** some CNEs are alignable between different paralogous genes in the same genome. After a segmental or whole-genome duplication, some paralogs retain some of their essential *cis*-regulatory elements in close proximity (McEwen et al., 2006; Woolfe and Elgar, 2007; Li et al., 2009; Tsang et al., 2009), which are very likely to represent enhancers. But even in fish that have undergone an additional whole-genome duplication, these "duplicated conserved non-coding elements" (dCNEs) are quite rare (~124, listed by McEwen et al. (2006)).
- **Length/Distances:** The length of vertebrate CNEs is better conserved than the length of conserved exons (Retelska et al., 2007). The distances between vertebrate CNEs (Sun et al., 2006) are better conserved than the distances between genes or between exons.
- **Position relative to genes:** around genes, vertebrate CNEs are evenly distributed between the 5′ and 3′ end. The regions farther away from genes are denser in conserved elements (Blanchette et al., 2006), 12% of duplicated CNEs (for these, a target gene is clearly assignable) are located farer than 1 MB from their target (Vavouri et al., 2006). CNEs are four times more common in "gene deserts," defined as >640 kb without a protein-coding gene, making up 25% of the human genome (Ovcharenko et al., 2005; Siepel et al., 2005). The longest of these regions flank some well-known developmental regulators like Otx2, Dach, Sall1 or Sox2 (see also Table 1).
- **Genome distribution:** the location of CNEs in the genome is not random. In flies flanking genes show over-representation of ion channel and cytoskeleton functions (Glazov et al., 2005; Papatsenko et al., 2006). In vertebrates, flies and worms, CNEs are associated with transcription factors (Sandelin et al., 2004) and under-represented around housekeeping genes (Farré et al., 2007; Vavouri et al., 2007). The initially reported bias towards the nervous system (Bejerano et al., 2004) seems to have been a statistical artifact (Taher and Ovcharenko, 2009).

chromatin structure (De Santa et al., 2010) and that many are bound by the transcriptional co-activator BCP (Kim et al., 2010). *In vivo* validated enhancers in the interleukin and IRX loci were shown to be transcribed as well (Jones and Flavell, 2005; de la Calle-Mustienes et al., 2005). RNA molecules of two transcribed enhancers even recruit transcription factors that then bind to the DNA sequences of the enhancers (Feng etal., 2006; Sanchez-Elsner et al., 2006; Bond et al., 2009). Although transcription seems to play a big role in *cis*-regulation in some individual examples, we are not aware of a general mechanism of how these RNA molecules are linked to regulatory sequences. Kim et al. (2010) note that transcription might be just a byproduct of open chromatin or one way to keep the chromatin accessible for transcription factors. Currently, the major lesson from these experiments is that experimenters should not remove transcribed sequences from their candidate elements when preparing an enhancer screen.

Apart from non-coding RNAs, another explanation for the high conservation of long sequences is the overlap of neighboring binding sites. An elegant example of this has been found in the enhancer of interferon-β. Several transcription factors bind to this 50 bp enhancer and form a complex called "enhanceosome" (Thanos and Maniatis, 1995). Panne et al. (2007) have combined several crystal structures to obtain one single 3D structure of the complex. Their model shows a general absence of protein interactions but instead a strong overlap of the different binding sites (Fig. 2). Such dense chains of proteins, contacting every single base pair of the DNA, might be one explanation for the high conservation of enhancers. In the context of enhancer analysis, the resulting interdependence between dozens of different binding sites within enhancers can make it difficult to predict the effect of targeted nucleotide mutations, once an element has been successfully identified in a screen. Therefore, researchers would be well advised to limit analysis on the binding site level to only a well-known candidate transcription factor.

Sequence conservation in chromatin-based assays

Despite the literature presented in the preceding sections, some fragments have been also shown to direct expression although they were not conserved: in the RET locus (Fisher et al., 2006), validated enhancers were not alignable between mammals and fish. In the case of PHOX2 (McGaughey et al., 2008), mammal/fish alignments fail to detect 40–70% of functional elements. Heart enhancers bound by P300 were a lot less conserved than enhancers active in the forebrain (Blow et al., 2010). Similar observations from *even-skipped* in flies and endo16 in sea urchin (Hare et al., 2008a,b; Wratten et al., 2006; Romano and Wray, 2003) showed that *cis*-regulatory blocks are not always detectable with standard local alignment tools when the evolutionary distance becomes

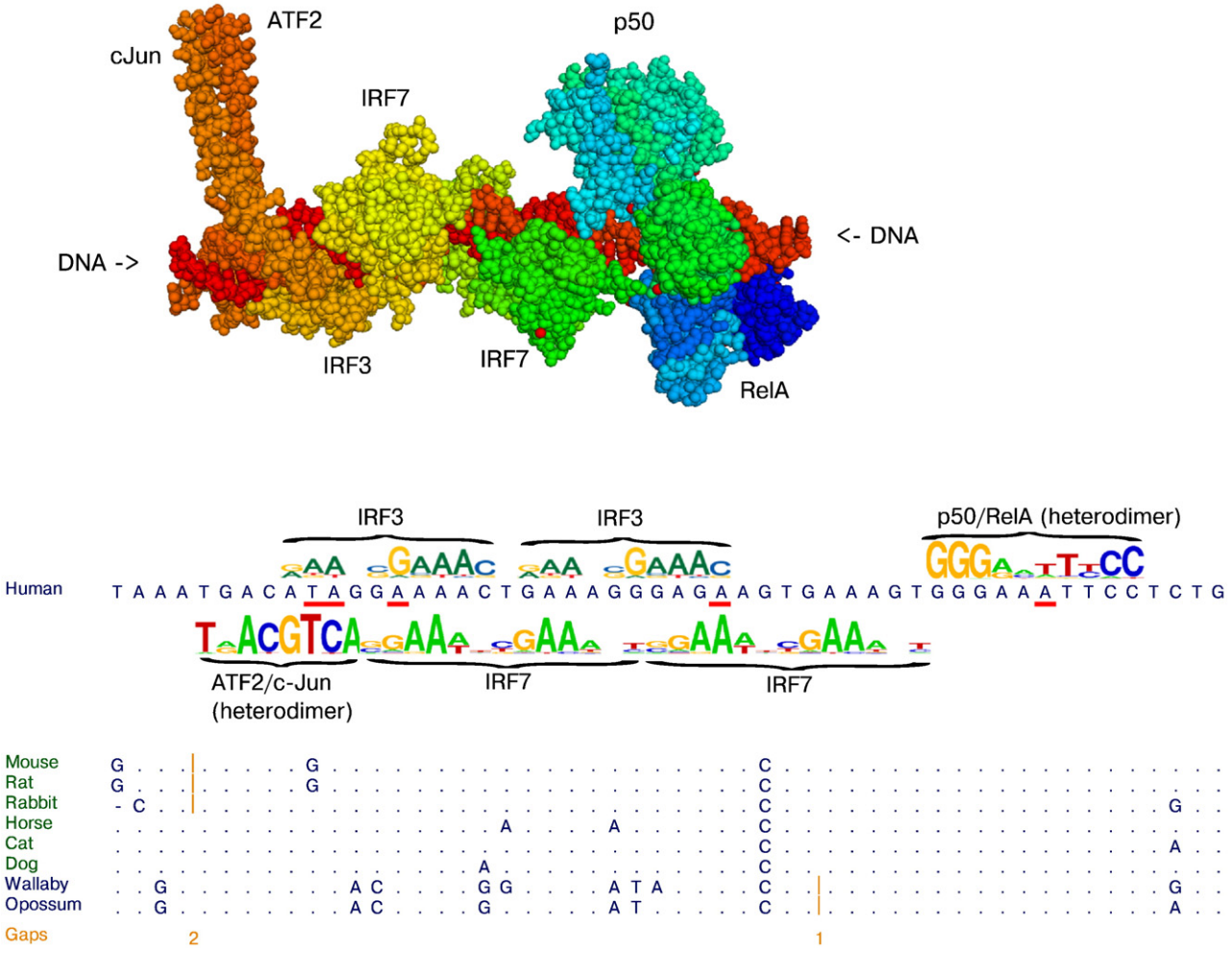

Fig. 2. Interferon enhanceosome: Crystal structure from Panne et al. (2007) and protein binding sites overlaid onto a multi-species alignment from the UCSC genome browser. Every base pair is contacted by at least one protein. Most base pairs that change in the multi-species alignment are at positions tolerated by the transcription factor binding properties. Positions where the nucleotides in the enhancer differ from the consensus of the transcription factor matrices are underlined in red. Due to these differences, some binding sites in this enhancer do not match the binding profile very well. At default settings, software predictions on this sequence are not very precise. Transfac Match 2010.1 predicts 97 sites on the enhancer of which only two correspond to the validated ones shown here. UniProbe predicts around 30 putative binding sites for various other transcription factors on this human enhancer sequence at default settings but not a single one corresponds to the expected sites. Matrices shown were obtained from Transfac (V$CREBP1CJUN_01, V$IRF7_01, V$NFKAPPAB50_01) (Matys et al., 2003) and UniProbe (IRF3) (Newburger and Bulyk, 2009) and converted to motif logos with STAMP (Mahony and Benos, 2007).

too large (reviewed by Nóbrega and Pennacchio, 2004). This might be due to extensive turnover of the binding sites (Oda-Ishii et al., 2005; Hare et al., 2008b). Adapting alignment algorithms to this problem is a topic of ongoing research (Gordân et al., 2010; Hu et al., 2008; Satija etal., 2008; Kantorovitz et al., 2007). Similar relaxed alignment criteria predicted several elements in Amphioxus which were successfully validated (Hufton et al., 2009).

A completely different set of experiments, the ENCODE project, could not find a link between conservation and enhancer function either. As an extension of the human genome project, ENCODE aimed at the identification of "functional elements in 1% of the human genome" (Birney et al., 2007). The data was mostly obtained with chromatin immunoprecipitation assays which promise to identify *cis*-regulatory elements much faster than *in vivo* injections. Subsequent computational analysis of the resulting fragments that were considered functional showed that they were not significantly enriched in regions under constraint in cross-species non-coding alignments (King et al., 2007; Zhang et al., 2007). This seems surprising in the light of results like Table 2 which concentrated with success onto conserved sequences.

Several factors might explain the differences between *in vivo* and *in vitro* assays: first, the transcription factors targeted by antibodies in ENCODE were mostly ubiquitous, like Sp1, Pol4, Taf1, and not tissue-specific. When antibodies are used against P300, a factor assumed to bind to tissue-specific elements, and cells are dissected out from mice, an enrichment of conserved sequence was indeed found (Visel et al., 2009b). More recent high-throughput ChIP experiments define enhancers based on histone modifications (Heintzman et al., 2009) or nucleosome dynamics (He et al., 2010). They correspond well to conserved elements, and predict tissue-specific elements better. Second, a region might be bound by a single factor but this does not necessarily reflect a function which is under selective pressure (Li etal., 2008). Chromatin studies predict function rather than proving it and their results still need to be confirmed by *in vivo* tests. Third, selective pressure seems to vary a lot depending on the function of the regulated gene (King et al., 2007) and of the element itself. A project of the size of ENCODE had to use widely available immortalized cell lines like HeLa and HL60, so a signal biased towards developmental regulators might be invisible on a whole-genome level.

Nuclear extracts from immortalized cell cultures are the main input material for the four experimental techniques that predict enhancers. The first and oldest one is DNaseI digestion for the detection of nuclease hypersensitive sites (Gross and Garrard, 1988). The second one is chromatin immunoprecipitation to pull down regions bound by antibodies against modified histones (Heintzman etal., 2009) or transcription factors. FAIRE (formaldehyde-assisted isolation of regulatory elements) exploits the fact that after formaldehyde-fixation and phenol-chloroform extraction, DNA in the aqueous phase is highly enriched for fragments in an open chromatin state. Chromatin conformation capture (3C) uses proximity ligation to quantify contacts between DNA regions (Dekker et al., 2002; Dostie et al., 2006). Whereas it might not predict enhancers per se, sequences that are found by 3C to interact with a promoter are likely to show a *cis*-regulatory function (Gheldof et al., 2010).

The common problem of these techniques is that results obtained from cell culture assays do not seem to expose tissue-specific elements (Attanasio et al., 2008; Göttgens et al., 2000). Many are binding the transcription factors but might still be nonfunctional (Li et al., 2008). Some enhancers predicted with cell cultures can become repressors in an *in vivo* context (Voth et al., 2009). Replacing cultured cells with ones manually dissected from animals can remedy these problems, but the feasibility of this approach depends on the size of the tissue: Heintzman et al. (2009) and Soshnikova and Duboule (2009) were obliged to sacrifice 150 and 75 mouse embryos, respectively, per experiment. The alternative, automated cell sorting requires an already known *cis*-regulatory element to mark the cells with fluorescence, e.g. blood or neurons (Long et al., 1997; Cerda et al., 2009; Jiang et al., 2008). Both approaches still depend on big amounts of nuclear extract, on the order of 10^7 to 10^8 cells, a problem that will soon become less critical with recent technical improvements of the immunoprecipitation procedure (Dahl and Collas, 2008), the ongoing replacement of microarrays with DNA sequencing (Wederell et al., 2008) and especially single-molecule sequencers (Goren et al., 2010). However, for most laboratories and with current protocols, although they certainly represent the future of *cis*-regulatory analysis, high-throughput assays are still difficult to apply to the limited number of cells that are typically found in developing embryos or tissue sub-structures.

Redundancy of regulatory elements

Expression patterns of enhancers in a single locus often seem to overlap (see Table 1). Hong et al. (2008) recently coined the term "shadow enhancer" for this phenomenon. They reason that the resulting redundancy protects essential developmental processes against mutations. However, while redundancy is often observable at early developmental stages, we are not aware of two enhancers with an identical pattern at all stages and the same quantitative expression level. Redundancy might explain why no phenotypic effect was observable in a laboratory environment when long stretches of non-coding sequences with several mega base pairs, including highly conserved elements or previously characterized enhancers for beta-globin, Engrailed2, Fgf4, Gata1 and Myod were knocked out in mice (Navas et al., 2006; Visel et al., 2007; Nóbrega et al., 2004; Ahituv et al., 2007a,b; Li Song and Joyner, 2000; Iwahori et al., 2004; Guyot et al., 2004; Chen and Goldhamer, 2004). But other researchers have observed the complete contrary: directed mutations of tissue-specific elements have lead to clear phenotypic effects in the loci of Shh, Shox, Meis1, Hoxc8, Dhand2 and Bmp2 (Lettice et al., 2003; Sabherwal et al., 2007; Xiong et al., 2009; Juan and Ruddle, 2003; Yanagisawa et al., 2003; Dathe et al., 2009), even when they involved just single base pairs as in the case of Shh or gamma-globin (Papachatzopoulou et al., 2007; Lettice et al., 2008; Rahimov et al., 2008). And two elements have to be deleted in combination to produce a visible phenotype in the locus of TCR-gamma (Xiong et al., 2002).

Taken together, the redundancy of regulatory elements resembles the redundancy of genes. It brings to mind a controversy on the exact function of HOX paralogs that started 15 years ago. Several of them were knocked out, some in combination, with the conclusion that redundancy is apparent in some tissues, some genes, and not in others (Horan et al., 1995; Condie and Capecchi, 1994). Therefore, partial co-activity of essential *cis*-regulatory sequences is expected for many essential processes, just like in genes. For a screen of putative elements, this increases the chance of the experimenter to find activating sequences in the tissue of interest but can render analysis by deletion (knock-out in genome or BACs) difficult to interpret in some cases.

Predicting tissue-specificity from nucleotide sequences

Distinctive features of enhancers

The guidelines from Box 2 should maximize the number of positive enhancers from a screen but they cannot select elements that are specific for a certain tissue. To tackle this question, one has to start to search for individual binding sites within the conserved elements, to find a link between the sequence and the function of enhancers. The basic idea to use the nucleotide sequence of *cis*-regulatory elements to predict the activating tissues is not new (Fondrat and Kalogeropoulos, 1994). Although the number of characterized sequences bound by transcription factors has exploded recently thanks to new techniques (Noyes et al., 2008; Badis et al., 2009; Newburger and Bulyk, 2009), the scheme's success depends on the detection of functional binding sites in the genome, a complex topic which has been reviewed elsewhere in

Box 2

Guidelines for enhancer screens:

- Confirm that *in vivo* screening is necessary: BAC/Cosmid constructs often reproduce the full expression pattern of a gene. Cell culture models are amenable to high-throughput assays and fast testing of individual elements.
- Invertebrate model organisms are the cheapest organisms to manipulate but their sequences cannot be mapped to vertebrates with current alignment algorithms. Assaying fragments in fish instead of mice can accelerate the assays. In some cases, non-coding alignments between paralogs, cloning DNA from close organisms with smaller genomes and injection of raw PCR fragments can simplify the experiments.
- A long proximal upstream region (5–7 kb) should be tested first, it could also give rise to an endogenous promoter for subsequent tests of enhancers. The TSS of a gene should be determined with care, taking into account all available cDNA/EST evidence or external databases like DBTSS (Wakaguri et al., 2008).
- Non-proximal elements should be tested with the endogenous promoter if possible. Otherwise, a "Super Promoter" is only available in flies. For vertebrates, (Gehrig et al., 2009) describe the best promoters depending on the tissue of interest.
- CNEs (conserved non-coding elements) should be preferred and can be located up to 1 MB away in mammals, skipping neighboring genes. The synteny of the locus can be taken into account when selecting them.
- CNEs with a conservation across the longest phylogenetic distance should be tested first and transcribed sequences are not to be excluded. Some classes of genes like transcription factors are flanked by more and better conserved elements, so the "best" phylogenetic distance depends on the gene of interest, it can be human/chicken in one case (Uchikawa et al., 2003), fish/human in another (Shin et al., 2005) or the best-conserved primate alignments (Prabhakar et al., 2006).
- The number of proteins binding to a conserved *cis*-regulatory element should not be underestimated.
- Unless the cell type is identical, one should be cautious when filtering elements based on large-scale chromatin data from cell cultures although more and more of them are becoming available.
- Sequence-based predictions heavily rely on the available data about the tissue of interest. They should be taken with a grain of salt if they make general assumptions on the composition of *cis*-regulatory elements. They can be tested on control gene sets (e.g. derived from microarrays, *in situ* screen databases or manually curated).
- Partial redundancy is expected and negative elements can be further characterized by combining them with others, as they might repress or modulate the activity of others.

detail (Wasserman and Sandelin, 2004; Vavouri and Elgar, 2005; Elnitski et al., 2006). The main difficulty here is that a degenerate motif equivalent to 4–6 base pairs (Maston et al., 2006) occurs very frequently in any long nucleotide sequence. This leads to the "futility theorem" which states that "essentially all predicted TFBS will have no functional role" in the cell (Wasserman and Sandelin, 2004) (see also Fig. 2), although the purified protein domain often binds the predicted oligonucleotides in gel shifts (Tronche et al., 1997). Therefore, in order to discriminate functionally valid sites from spurious sequence matches, several additional features have been proposed.

One of them is helical spacing, which leads to preferred distances between sites, as a complete turn of the DNA takes about 10 base pairs: this is clearly supported by experimental data for the *bicoid* transcription factor (Hanes et al., 1994), visible as periodic signals when searching known enhancers for its motif (Makeev et al., 2003). Papatsenko et al. (2009) also found helical phasing limited to certain transcription factors, again *bicoid* and to a lesser extent, *caudal*/*distalless*. Nevertheless, Berman et al. (2004), for instance, did not observe helical spacing in a list of *Drosophila* enhancers, and Li et al. (2007) found them only in blastoderm stage enhancers. In yeast, similar observations have been corrected recently (Yuan et al., 2007), noting a very weak link between inter-site distances and the expression pattern.

The second criterion involves the strength of the match: as transcription factors recognize degenerate sequences, sites can correspond more or less to the consensus. In the case of Su(H), the factor recognized mostly optimal consensus sequences (Adryan et al., 2007). In the case of Foxa and Rest (Gaudet and Mango, 2002; Bruce et al., 2009), the affinity of the site to the protein seems to indicate that enhancers are active. A compelling evidence in yeast is that binding site fuzziness increases with the density of sites in a promoter region (Bilu and Barkai, 2005). However, in some high-throughput assays, regions that lack the E2F consensus motif can be bound very well (Rabinovich et al., 2008). Similarly, in the interferon-β enhanceosome (Panne et al., 2007) mentioned before, proteins are bound to very weak sites (see Fig. 2), and a model of *Drosophila* patterning highlights the role of weak sites as well (Segal et al., 2008). Overall, the conflicting evidence makes it hard to give advice on the optimum affinity of factors when searching for functional binding sites.

The third proposed property is "homotypic clustering," binding sites that occur in several, possibly degenerate, copies. As transcription factors track along the chromatin (Gorman and Greene, 2008), this is thought to increase the thermodynamic probability of binding to a site. Although several studies on Su(H) (Adryan et al., 2007), Stat5 (Pena and Whitelaw, 2005) and *C. elegans* interneuron enhancers (Wenick and Hobert, 2004) observed that just a single binding site with no additional copies was sufficient for expression, it seems that homotypic clustering is a general feature of enhancers involved in fly blastoderm patterning (Rebeiz et al., 2002; Lifanov et al., 2003; Segal et al., 2008) but not in other tissues (Li et al., 2007). A filter based on this criterion led to the identification of new enhancers when searching the *Drosophila* genome (Berman et al., 2002; Markstein et al., 2002) and in mammalian genomes, it led to non-random predictions in a whole-genome scan for predicted binding sites for factors like p53 or Rest (Zhang et al., 2006). Homotypic clusters are 25-fold enriched in developmental enhancers, an indication that homotypic clustering might depend on the types of factors (Gotea et al., 2010). This suggests a situation like in yeast, where only some types of binding sites tend to occur in homotypic clusters (Harbison et al., 2004). Therefore, more recent algorithms favor thermodynamical models that evaluate all matches in a certain window. In such a scheme, a homotypic cluster of several weak copies and a single strong match can obtain the same score (Gertz et al., 2009; Roider et al., 2007).

General approach of the algorithms

Based on the rules listed in the previous chapter or on similar rules, some software algorithms claim to detect all sequences with any *cis*-regulatory potential in the genome (Pierstorff et al., 2006; Taylor et al.,

2006). The idea here is that all *cis*-elements, independent of the tissue where they are active and its transcription factors, share certain nucleotide characteristics. However, softwares can only be tested on a limited set of enhancers, from a certain type of experiment, so the results risk being biased towards the tissues on which the models were trained on. *Drosophila* is the only animal where benchmarks and systematic comparisons of such algorithms currently exist. But even when motif sets were tailored to a specific tissue, performance varied widely, with many tissues being impossible to predict (Kantorovitz et al., 2009).

Altogether, evidence for homotypic clustering, spatial constraints or protein affinity seems to depend on the type of transcription factor analyzed. It is therefore difficult to find a general rule that distinguishes functional from spurious binding site matches which would be valid for all factors, tissues and organisms.

On the other hand, some rules have been found in examples from certain tissues. We selected studies that predicted tissue-specific enhancers and then tested the resulting DNA fragments afterwards (see Supplementary Table 1). These approaches share a common setup: the starting point is either a collection of previously described and co-expressed enhancers from which common motifs are extracted *de novo*, without any knowledge of the factors that bind to them (for reviews on this step see Sandve and Drabløs, 2006; Tompa et al., 2005; MacIsaac and Fraenkel, 2006). The alternative to motif inference from sequences is a set of well-known tissue-specific transcription factors and their DNA-specificities, like *Dorsal* in the case of dorsal–ventral patterning. The newly discovered or already known short DNA motifs are then used to search the genome or around some loci of interest for similar sequences. The crucial part is to define the "similarity" of a sequence, in the absence of BLAST-statistics that require longer alignments. Do two weak matches score higher than one strong match? How many binding sites are necessary to trigger a match? Does one (longer) *Dorsal* site score as well as two (shorter) *Twist* sites? Researchers have answered these questions very differently, varying with the organism and tissue of interest, taking into account some of the general properties discussed above.

The simplest score was the number of exact binding site matches within a certain window size, e.g. three *Dorsal* binding sites within 400 bp (Markstein et al., 2002) or two conserved OTX sites within 125 bp (Haeussler et al., 2010). The most complex approach took into account the affinity of the DNA sequences to the transcription factor, competition between sites, helical spacing and the order of conservation in a second species (Hallikas et al., 2006). In all cases, the sequences are scanned according to this model, regions that exceed a minimum score are reported and annotated when they are already known from the literature.

Benchmarking predictions

Previously uncharacterized predictions can be evaluated in two ways: researchers can either determine the expression patterns of the flanking genes or test the predicted enhancers themselves with a reporter gene. Like all predictions, these are unlikely to achieve 100% accuracy. The most interesting performance measure in the context of an enhancer screen is the enrichment relative to a background: if 30% of all genes in the organism are expressed in a tissue (background or random rate) but the positive share increases to 60% among the predictions, then this corresponds to a two-fold enrichment. In Supplementary Table 1, we show this measure and also added binomial *p*-values to indicate how significant the difference to the background is, taking into account the number of experiments. Obviously, for enhancer tests with a reporter gene, background rate and *p*-values are difficult to determine, as the total number of active enhancers for a given tissue is not known in any organism.

What can we conclude from the studies summarized in Supplementary Table 1? First of all, the majority focus on invertebrate model organisms. The reasons are certainly experimental advantages, like compact upstream sequences fitting into a single plasmid and developmental times measured in days. This is especially relevant when testing predictions, as enough sequences have to be assayed to show statistical significance. Wang et al. (2006) based their prediction on only one site of one transcription factor (GATA1), whose expression in blood cell precursors had been shown to directly lead to their terminal differentiation. Furthermore, most approaches have focused on two examples: *Drosophila* blastoderm patterning and mammalian muscle cells. The latter is one of the best-described models of transcriptional regulation in animals, with many characterized enhancers as training data. For muscle cells, an abundant literature has been published and one of the first and most-cited enhancer sequence analyses summarized this as an annotated control data set (Wasserman and Fickett, 1998). In both cases, upstream transcription factors had been identified by previous studies, their binding sites could be searched and validated against the known data which in turn motivated further experimental validations. Therefore, the only algorithm (Schroeder et al., 2004) where all predicted fragments were really enhancers, could build on decades of research on *Drosophila* patterning and searched for known binding sites of nine well-known transcription factors. We note that the complexity of a prediction algorithm seems to be less important than the type of the cells and previous knowledge about them: one of the highest rates of correctly predicted genes is achieved by a straightforward single-motif search, based on genes expressed in only two individual interneuron cells in *C. elegans*. Muscle gene identification with several previously completely uncharacterized motifs leads to merely a 2-fold enrichment. A duplicated motif search for anterior nervous system enhancers obtained a 3-fold enrichment. Both are still some improvement compared to random selection. In the end, the decision to use the algorithms will depend whether enrichment values of 2–4 appear high enough to justify the risk of missing an enhancer. We summarize general advice of nucleotide-base enhancer-tissue predictions in Box 3 and the possible steps in an enhancer screen as a workflow in Fig. 3.

Perspectives

In this review, we have presented the different possible ways to identify functional, tissue-specific enhancers in a given locus. Nonfunctional elements attract limited interest, as negative results often do not encourage further study. Elements considered nonfunctional should nevertheless be documented, since they can be important as negative controls in subsequent *cis*-regulatory screening or modeling. Many of the elements without activity may contain tissue-specific silencers, a topic too often neglected in the field. Very little is known about their difference from activating elements or whether there are any constraints on the distance to their targets. Testing some of the putative silencers from Table 2 in combination with a well-known enhancer could be a first step. As quantitative differences are difficult to measure with GFP and LacZ reporter genes, an *in vivo* luciferase assay like in (Shin et al., 2005) might be appropriate, and was also used for the characterization of the Endo16 amplifier effect in sea urchin (Yuh et al., 1998).

On the computational side, some of the software tools make searches for short motifs in conserved *cis*-regulatory elements applicable on whole mammalian genomes. However, it is surprisingly difficult to link the resulting matches with the already known gene data. Simple tasks like the annotation of flanking genes still require programming and the extraction of tissue-specific genes from *in situ* databases is far from trivial. In addition, programs like EEL, Ahab and Clusterdraw allow scanning only one set of motifs at a time, mandating a "trial and error" approach (Palin et al., 2006), although a control data set of tissue-specific genes would permit automatic optimization of all parameters.

Both computational algorithms and wet-lab users would benefit from better curation of published studies. They first need training and

Box 3

Enhancer prediction based on short sequence motifs

Most studies confirmed that the tissue-specific factors do leave a trace in the non-coding sequences of the gene they regulate. But they do not allow to point out a clearly superior search algorithm as the particular benchmark sets and cell types have little in common. Ahab (Rajewsky et al., 2002) and the similar but faster Cluster–Draw (Papatsenko, 2007) based on thermodynamical foundations obtain convincing results in the case of *Drosophila* patterning and the programs are available and easy to run. But they do not take into account conserved regions. EEL is the only program that focuses on conserved regions, has been validated in experiments and can be run on any computer (Palin et al., 2006). It is also the only one based on the assumption that binding site order has to be conserved, for which there is not a lot support in the literature, to our knowledge.

Detailed protocols on the practical application of enhancer prediction tools have been published elsewhere (Smith, 2008; Papatsenko and Levine, 2005; Palin et al., 2006). Most of these tools have been trained on muscle or blastoderm patterning but can be easily applied on genes that are not expressed in these cell types. They promise to reduce the number of *in vivo* tests by filtering out sequences that do not fit to the model. Before validating these predictions with experiments, one should consider other ways to benchmark the predictions. Do the motifs used correspond to known key transcription factors in the tissue of interest? Is there a control set of known enhancers, perhaps extractable from the literature, similar to the ones from *Drosophila* from (Halfon et al., 2008)? If there is none, can predictions be assessed by checking the expression pattern of the flanking genes (Papatsenko and Levine, 2005), potentially by using an *in situ* database such as the ones listed in (Armit, 2007)?

benchmarking data to tune their algorithms. The latter have difficulty finding already validated enhancers that drive in the right tissue although they might have been already isolated in a different locus and for different purposes, but lack the necessary keywords in the abstract. Although more and more *cis*-regulatory analyses are available, vertebrate model organism databases currently do not curate transgenic sequences at all (MGI) or just indicate expression patterns for some of them (Zfin) from publications. General databases like Genbank or third-party projects like ORegAnno (Griffith et al., 2008) store sequences but not the expression patterns, as they lack the species-specific anatomical knowledge. As a result, there is no database yet where one can find a comprehensive list with sequences of already characterized enhancers in mice that are expressed in a given tissue. It is in the interest of the scientific community working on vertebrates that model organism databases start to annotate sequences and expression patterns of enhancers from the literature, as it is current practice in invertebrates like *Drosophila* (Halfon et al., 2008; Ivan et al., 2008), *C. intestinalis* (Sierro et al., 2006; Tassy et al., 2006) or *C. elegans* (Lee, 2005).

Given the explosion of high-throughput data from chromatin immunoprecipitation and similar assays, it is ever more important to compare them thoroughly with low-throughput *in vivo* results. This includes the negative sequences, as controls for future assays. With more well-characterized and curated enhancers, cell-specific motif signatures should emerge that will help to identify similar *cis*-regulatory elements based on their sequences, a technique which is currently efficient only in *Drosophila* blastoderm patterning (Segal et al., 2008; Zinzen et al., 2009; Kantorovitz et al., 2009).

Just how many such signatures will be needed is open to speculation. If non-coding sequences are conserved due to overlapping transcription factors, then the density of binding sites in the conserved part of genome should be at a similar level as in the

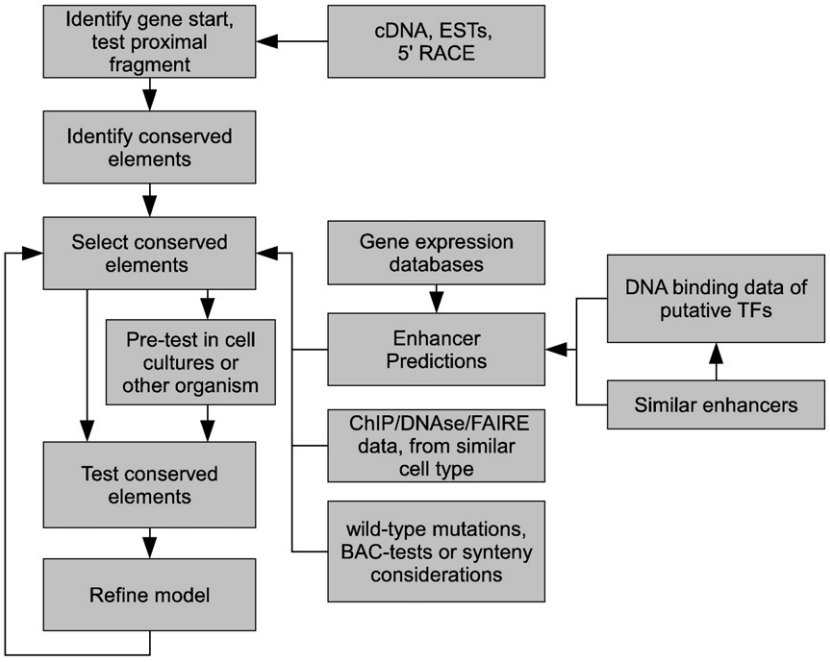

Fig. 3. An overview of the different possible steps during an enhancer screen. This corresponds to the approach of current model-based enhancer predictions like those presented by Segal et al. (2008) or Kantorovitz et al. (2009).

interferon-β enhanceosome. A sequence that is conserved with mouse over 500 bp could then be bound by 100 proteins, a level of complexity different from the estimation based on experimental data, a decade ago, that around five different factors were binding per *cis*-regulatory element (Arnone and Davidson, 1997). A high density of binding sites would also better fit the results of ChIP assays that predict tens of thousands of sites per factor (Marson et al., 2008; Margolin et al., 2009).

The tissue-specificity of many regulatory elements and possible interactions during development combined with the high price of *in vivo* testing makes it difficult to imagine that all functional sequences in an animal genome will ever be completely identified with any current technology. It will certainly take many new benchmark collections, comprehensive modeling softwares and new experimental screens to explain why and how enhancers activate their target genes in a given cell population.

Supplementary materials related to this article can be found online at doi:10.1016/j.ydbio.2010.11.026.

Acknowledgments

We wish to thank Yan Jaszczyszyn, Patrick Lemaire and Sylvie Rétaux for the careful rereading of the manuscript. Our work was supported by INRA and CNRS, the French GIS Institut de la Génomique Marine, the Marine Genomics Network of Excellence (EU-FP6 contract no. GOCE-CT-2004-505403), the ANR projects CHOREGNET and CHOREVONET, and the Plurigenes STREP project LSHG-CT-2005-018673. M.H. received funding from a Marie Curie Early Stage Research Training Fellowship (MEST-CT-2004-504854) and the Plurigenes STREP project LSHG-CT-2005-018673.

References

Adryan, B., Woerfel, G., Birch-Machin, I., et al., 2007. Genomic mapping of suppressor of hairy-wing binding sites in Drosophila. Genome Biol. 8, R167.
Ahituv, N., Akiyama, J., Chapman-Helleboid, A., et al., 2007a. In vivo characterization of human Apoa5 haplotypes. Genomics 90, 674–679.
Ahituv, N., Prabhakar, S., Poulin, F., et al., 2005. Mapping cis-regulatory domains in the human genome using multi-species conservation of synteny. Hum. Mol. Genet. 14, 3057–3063.
Ahituv, N., Zhu, Y., Visel, A., et al., 2007b. Deletion of ultraconserved elements yields viable mice. PLoS Biol. 5, e234.
Akbari, O.S., Bousum, A., Bae, E., et al., 2006. Unraveling cis-regulatory mechanisms at the abdominal-a and abdominal-b genes in the drosophila bithorax complex. Dev. Biol. 293, 294–304.
Amaral, P.P, Mattick, J.S., 2008. Noncoding RNA in development. Mamm. Genome 19, 454–492.
Aparicio, S., Morrison, A., Gould, A., et al., 1995. Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, *Fugu rubripes*. Proc. Natl. Acad. Sci. U.S.A. 92, 1684–1688.
Arányi, T., Faucheux, B.A., Khalfallah, O., et al., 2005. The tissue-specific methylation of the human tyrosine hydroxylase gene reveals new regulatory elements in the first exon. J. Neurochem. 94, 129–139.
Armit, C., 2007. Developmental biology and databases: how to archive, find and query gene expression patterns using the world wide web. Organogenesis 3, 70–73.
Arnone, M.I., Davidson, E.H., 1997. The hardwiring of development: organization and function of genomic regulatory systems. Development (Cambridge, England) 124, 1851–1864.
Arnosti, D.N., 2003. Analysis and function of transcriptional regulatory elements: insights from Drosophila. Annu. Rev. Entomol. 48, 579–602.
Attanasio, C., Reymond, A., Humbert, R., et al., 2008. Assaying the regulatory potential of mammalian conserved non-coding sequences in human cells. Genome Biol. 9, R168.
Badis, G., Berger, M.F., Philippakis, A.A., Talukder, S., Gehrke, A.R., Jaeger, S.A., Chan, E.T., Metzler, G., Vedenko, A., Chen, X., Kuznetsov, H., Wang, C., Coburn, D., Newburger, D.E., Morris, Q., Hughes, T.R., Bulyk, M.L., 2009. Diversity and complexity in DNA recognition by transcription factors. Science 324, 1720–1723.
Banerji, J., Rusconi, S., Schaffner, W., 1981. Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. Cell 27, 299–308.
Barthel, K.K.B., Liu, X., 2008. A transcriptional enhancer from the coding region of Adamts5. PLoS ONE 3, e2184.
Barton, L.M., Gottgens, B., Gering, M., et al., 2001. Regulation of the stem cell leukemia (PAX) gene: a tale of two fishes. Proc. Natl. Acad. Sci. USA 98, 6747–6752.
Bejerano, G., Lowe, C.B., Ahituv, N., et al., 2006. A distal enhancer and an ultraconserved exon are derived from a novel retroposon. Nature 441, 87–90.
Bejerano, G., Pheasant, M., Makunin, I., et al., 2004. Ultraconserved elements in the human genome. Science 304, 1321–1325.
Bergman, C.M., Kreitman, M., 2001. Analysis of conserved noncoding DNA in Drosophila reveals similar constraints in intergenic and intronic sequences. Genome Res. 11, 1335–1345.
Berman, B.P., Nibu, Y., Pfeiffer, B.D., et al., 2002. Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. Proc. Natl. Acad. Sci. USA 99, 757–762.
Berman, B.P., Pfeiffer, B.D., Laverty, T.R., et al., 2004. Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in Drosophila melanogaster and Drosophila pseudoobscura. Genome Biol. 5, R61.
Bessa, J., Tena, J.J., de la Calle-Mustienes, E., Fernández-Miñán, A., Naranjo, S., Fernández, A., Montoliu, L., Akalin, A., Lenhard, B., Casares, F., Gómez-Skarmeta, J.L., 2009. Zebrafish enhancer detection (ZED) vector: a new tool to facilitate transgenesis and the functional analysis of cis-regulatory regions in zebrafish. Dev. Dyn. Off. Publ. Am. Assoc. Anatomists 238, 2409–2417.
Bilu, Y., Barkai, N., 2005. The design of transcription-factor binding sites is affected by combinatorial regulation. Genome Biol. 6, R103.
Birney, E., Stamatoyannopoulos, J.A., Dutta, A., et al., 2007. Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. Nature 447, 799–816.
Blanchette, M., Bataille, A.R., Chen, X., et al., 2006. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. Genome Res. 16, 656–668.
Blow, M.J., McCulley, D.J., Li, Z., et al., 2010. Chip-seq identification of weakly conserved heart enhancers. Nat. Genet. 42, 806–810.
Bond, A.M., Vangompel, M.J.W., Sametsky, E.A., Clark, M.F., Savage, J.C., Disterhoft, J.F., Kohtz, J.D., 2009. Balanced gene regulation by an embryonic brain ncRNA is critical for adult hippocampal GABA circuitry. Nat. Neurosci. 12, 1020–1027.
Bouchard, M., Grote, D., Craven, S.E., et al., 2005. Identification of Pax2-regulated genes by expression profiling of the mid-hindbrain organizer region. Development 132, 2633–2643.
Boulin, T., Etchberger, J.F., Hobert, O., 2006. Reporter gene fusions. WormBook 1–23.
Bruce, A.W., López-Contreras, A.J., Flicek, P., et al., 2009. Functional diversity for rest (nrsf) is defined by in vivo binding affinity hierarchies at the DNA sequence level. Genome Res.
Butler, J.E., Kadonaga, J.T., 2001. Enhancer–promoter specificity mediated by dpe or tata core promoter motifs. Genes Dev. 15, 2515–2519.
Cameron, R.A., Chow, S.H., Berney, K., et al., 2005. An evolutionary constraint: strongly disfavored class of change in DNA sequence during divergence of cis-regulatory modules. Proc. Natl. Acad. Sci. USA 102, 11769–11774.
Cameron, R.A., Oliveri, P., Wyllie, J., et al., 2004. Cis-regulatory activity of randomly chosen genomic fragments from the sea urchin. Gene Expr. Patterns 4, 205–213.
Carter, D., Chakalova, L., Osborne, C.S., et al., 2002. Long-range chromatin regulatory interactions *in vivo*. Nat. Genet. 32, 623–626.
Casillas, S., Barbadilla, A., Bergman, C.M., 2007. Purifying selection maintains highly conserved noncoding sequences in Drosophila. Mol. Biol. Evol. 24, 2222–2234.
Cerda, G.A., Hargrave, M., Lewis, K.E., 2009. RNA profiling of FAC-sorted neurons from the developing zebrafish spinal cord. Dev. Dyn. 238, 150–161.
Chen, H., Blanchette, M., 2007. Detecting non-coding selective pressure in coding regions. BMC Evol. Biol. 7 (Suppl 1), S9.
Chen, J., Huisinga, K.L., Viering, M.M., et al., 2002. Enhancer action in trans is permitted throughout the Drosophila genome. Proc. Natl. Acad. Sci. USA 99, 3723–3728.
Chen, J.C.J., Goldhamer, D.J., 2004. The core enhancer is essential for proper timing of Myod activation in limb buds and branchial arches. Dev. Biol. 265, 502–512.
Chi, X., Chatterjee, P.K., Wilson III, W., Zhang, S., Demayo, F.J., Schwartz, R.J., 2005. Complex cardiac Nkx2-5 gene expression activated by noggin-sensitive enhancers followed by chamber-specific modules. Proc. Natl Acad. Sci. USA 102, 13490–13495.
Condie, B.G., Capecchi, M.R., 1994. Mice with targeted disruptions in the paralogous genes Hoxa-3 and Hoxd-3 reveal synergistic interactions. Nature 370, 304–307.
Conradt, B., Horvitz, H.R., 1999. The Tra-1a sex determination protein of c. elegans regulates sexually dimorphic cell deaths by repressing the Egl-1 cell death activator gene. Cell 98, 317–327.
Conte, I., Bovolenta, P., 2007. Comprehensive characterization of the cis-regulatory code responsible for the spatio-temporal expression of olSix3.2 in the developing medaka forebrain. Genome Biol. 8, R137.
Dahl, J.A., Collas, P., 2008. Microchip–a rapid micro chromatin immunoprecipitation assay for small cell samples and biopsies. Nucleic Acids Res. 36, e15.
Dathe, K., Kjaer, K.W., Brehm, A., et al., 2009. Duplications involving a conserved regulatory element downstream of BMP2 are associated with brachydactyly type a2. Am. J. Hum. Genet. 84, 483–492.
De Grassi, A., Segala, C., Iannelli, F., Volorio, S., Bertario, L., Radice, P., Bernard, L., Ciccarelli, F.D., 2010. Ultradeep sequencing of a human ultraconserved region reveals somatic and constitutional genomic instability. PLoS Biol. 8, e1000275.
de la Calle-Mustienes, E., Feijóo, C.G., Manzanares, M., et al., 2005. A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate iroquois cluster gene deserts. Genome Res. 15, 1061–1072.
De Val, S., Chi, N.C., Meadows, S.M., Minovitsky, S., Anderson, J.P., Harris, I.S., Ehlers, M.L., Agarwal, P., Visel, A., Xu, S., Pennacchio, L.A., Dubchak, I., Krieg, P.A., Stainier, D.Y.R., Black, B.L., 2008. Combinatorial regulation of endothelial gene expression by ETS and Forkhead transcription factors. Cell 135, 1053–1064.
Dekker, J., Rippe, K., Dekker, M., et al., 2002. Capturing chromosome conformation. Science 295, 1306–1311.
Denoeud, F., Kapranov, P., Ucla, C., et al., 2007. Prominent use of distal 5′ transcription start sites and discovery of a large number of additional exons in ENCODE regions. Genome Res. 17, 746–759.

Dermitzakis, E.T., Reymond, A., Scamuffa, N., et al., 2003. Evolutionary discrimination of mammalian conserved non-genic sequences (cngs). Science 302, 1033–1035.
Dong, X., Fredman, D., Lenhard, B., 2009a. Synorth: exploring the evolution of synteny and long-range regulatory interactions in vertebrate genomes. Genome Biol. 10, R86.
Dong, X., Navratilova, P., Fredman, D., Drivenes, O., Becker, T.S., Lenhard, B., 2009b. Exonic remnants of whole-genome duplication reveal cis-regulatory function of coding exons. Nucleic Acids Research.
Dorsett, D., 1993. Distance-independent inactivation of an enhancer by the suppressor of hairy-wing DNA-binding protein of Drosophila. Genetics 134, 1135–1144.
Dostie, J., Richmond, T.A., Arnaout, R.A., et al., 2006. Chromosome conformation capture carbon copy (5c): a massively parallel solution for mapping interactions between genomic elements. Genome Res. 16, 1299–1309.
Drake, J.A., Bird, C., Nemesh, J., et al., 2006. Conserved noncoding sequences are selectively constrained and not mutation cold spots. Nat. Genet. 38, 223–227.
Dubchak, I., Brudno, M., Loots, G.G., et al., 2000. Active conservation of noncoding sequences revealed by three-way species comparisons. Genome Res. 10, 1304–1306.
Ejsmont, R.K., Sarov, M., Winkler, S., et al., 2009. A toolkit for high-throughput, cross-species gene engineering in Drosophila. Nat. Methods.
Elnitski, L., Jin, V.X., Farnham, P.J., et al., 2006. Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. Genome Res. 16, 1455–1464.
Engström, P.G., Ho Sui, S.J., Drivenes, O., et al., 2007. Genomic regulatory blocks underlie extensive microsynteny conservation in insects. Genome Res. 17, 1898–1908.
Ertzer, R., Müller, F., Hadzhiev, Y., et al., 2007a. Cooperation of sonic hedgehog enhancers in midline expression. Dev. Biol. 301, 578–589.
Ertzer, R., Müller, F., Hadzhiev, Y., Rathnam, S., Fischer, N., Rastegar, S., Strähle, U., 2007b. Cooperation of sonic hedgehog enhancers in midline expression. Dev. Biol. 301, 578–589.
Farhadi, H.F., Lepage, P., Forghani, R., Friedman, H.C.H., Orfali, W., Jasmin, L., Miller, W., Hudson, T.J., Peterson, A.C., 2003. A combinatorial network of evolutionarily conserved myelin basic protein regulatory sequences confers distinct glial-specific phenotypes. J. Neurosci. Off. J. Soc. Neurosci. 23, 10214–10223.
Farré, D., Bellora, N., Mularoni, L., et al., 2007. Housekeeping genes tend to show reduced upstream sequence conservation. Genome Biol. 8, R140.
Feng, J., Bi, C., Clark, B.S., et al., 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.
Fisher, S., Grice, E.A., Vinton, R.M., et al., 2006. Conservation of RET regulatory function from human to zebrafish without sequence similarity. Science 312, 276–279.
Fondrat, C., Kalogeropoulos, A., 1994. Approaching the function of new genes by detection of their potential upstream activation sequences in *Saccharomyces cerevisiae*: application to chromosome iii. Curr. Genet. 25, 396–406.
Forrester, W.C., Epner, E., Driscoll, M.C., Enver, T., Brice, M., Papayannopoulou, T., Groudine, M., 1990. A deletion of the human beta-globin locus activation region causes a major alteration in chromatin structure and replication across the entire beta-globin locus. Genes Dev. 4, 1637–1649.
Frith, M.C., Valen, E., Krogh, A., et al., 2008. A code for transcription initiation in mammalian genomes. Genome Res. 18, 1–12.
Gaudet, J., Mango, S.E., 2002. Regulation of organogenesis by the *Caenorhabditis elegans* Foxa protein pha-4. Science 295, 821–825.
Gehrig, J., Reischl, M., Kalmár, E., Ferg, M., Hadzhiev, Y., Zaucker, A., Song, C., Schindler, S., Liebel, U., Müller, F., 2009. Automated high-throughput mapping of promoter-enhancer interactions in zebrafish embryos. Nat. Meth. 6, 911–916.
Gertz, J., Siggia, E.D., Cohen, B.A., 2009. Analysis of combinatorial cis-regulation in synthetic and genomic promoters. Nature 457, 215–218.
Geschwind, D., 2004. GENSAT: a genomic resource for neuroscience research. Lancet Neurol. 3, 82.
Ghanem, N., Jarinova, O., Amores, A., Long, Q., Hatch, G., Park, B.K., Rubenstein, J.L.R., Ekker, M., 2003. Regulatory roles of conserved intergenic domains in vertebrate Dlx bigene clusters. Genome Res. 13, 533–543.
Ghanem, N., Yu, M., Long, J., Hatch, G., Rubenstein, J.L.R., Ekker, M., 2007. Distinct cis-regulatory elements from the Dlx1/Dlx2 locus mark different progenitor cell populations in the ganglionic eminences and different subtypes of adult cortical interneurons. J. Neurosci. Off. J. Soc. Neurosci. 27, 5012–5022.
Gheldof, N., Smith, E.M., Tabuchi, T.M., et al., 2010. Cell-type-specific long-range looping interactions identify distant regulatory elements of the cftr gene. Nucleic Acids Res. 38, 4325–4336.
Glazov, E.A., Pheasant, M., McGraw, E.A., Bejerano, G., 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.
Goode, D.K., Snell, P., Smith, S.F., et al., 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.
Gordân, R., Narlikar, L., Hartemink, A.J., 2010. Finding regulatory DNA motifs using alignment-free evolutionary conservation information. Nucleic Acids Res.
Goren, A., Ozsolak, F., Shoresh, N., Ku, M., Adli, M., Hart, C., Gymrek, M., Zuk, O., Regev, A., Milos, P.M., Bernstein, B.E., 2010. Chromatin profiling by directly sequencing small quantities of immunoprecipitated DNA. Nat. Meth. 7, 47–49.
Gorman, J., Greene, E.C., 2008. Visualizing one-dimensional diffusion of proteins along DNA. Nat. Struct. Mol. Biol. 15, 768–774.
Gotea, V., Visel, A., Westlund, J.M., et al., 2010. Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. Genome Res. 20, 565–577.
Göttgens, B., Barton, L.M., Gilbert, J.G., et al., 2000. Analysis of vertebrate scl loci identifies conserved enhancers. Nat. Biotechnol. 18, 181–186.
Griffith, O.L., Montgomery, S.B., Bernier, B., et al., 2008. ORegAnno: an open-access community-driven resource for regulatory annotation. Nucleic Acids Res. 36, D107–D113.
Gross, D.S., Garrard, W.T., 1988. Nuclease hypersensitive sites in chromatin. Annu. Rev. Biochem. 57, 159–197.
Guerrero, L., Marco-Ferreres, R., Serrano, A.L., Arredondo, J.J., Cervera, M., 2010. Secondary enhancers synergise with primary enhancers to guarantee fine-tuned muscle gene expression. Dev. Biol. 337, 16–28.
Guo, G., Bauer, S., Hecht, J., et al., 2008. A short ultraconserved sequence drives transcription from an alternate Fbn1 promoter. Int. J. Biochem. Cell Biol. 40, 638–650.
Guyot, B., Valverde-Garduno, V., Porcher, C., et al., 2004. Deletion of the major Gata1 enhancer hs 1 does not affect eosinophil Gata1 expression and eosinophil differentiation. Blood 104, 89–91.
Haeussler, M., Jaszczyszyn, Y., Christiaen, L., Joly, J.S., 2010. A cis-regulatory signature for chordate anterior neurectodermal genes. PLoS Genet. 6, e1000912.
Halfon, M.S., Gallo, S.M., Bergman, C.M., 2008. Redfly 2.0: an integrated database of cis-regulatory modules and transcription factor binding sites in Drosophila. Nucleic Acids Res. 36, D594–D598.
Hanes, S.D., Riddihough, G., Ish-Horowicz, D., et al., 1994. Specific DNA recognition and intersite spacing are critical for action of the bicoid morphogen. Mol. Cell. Biol. 14, 3364–3375.
Hallikas, O., Palin, K., Sinjushina, N., et al., 2006. Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. Cell 124, 47–59.
Harbison, C.T., Gordon, D.B., Lee, T.I., et al., 2004. Transcriptional regulatory code of a eukaryotic genome. Nature 431, 99–104.
Hardison, R., Miller, W., 1993. Use of long sequence alignments to study the evolution and regulation of mammalian globin gene clusters. Mol. Biol. Evol. 10, 73–102.
Hardison, R.C., Oeltjen, J., Miller, W., 1997. Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. Genome Res. 7, 959–966.
Hare, E.E., Peterson, B.K., 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., et al., 2008b. Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. PLoS Genet. 4, e1000106.
He, H.H., Meyer, C.A., Shin, H., et al., 2010. Nucleosome dynamics define transcriptional enhancers. Nat. Genet. 42, 343–347.
Heintzman, N.D., Hon, G.C., Hawkins, R.D., et al., 2009. Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature 459, 108–112.
Hobert, O., 2002. PCR fusion-based approach to create reporter gene constructs for expression analysis in transgenic C. elegans. Biotechniques 32, 728–730.
Hobert, O., 2008. Regulatory logic of neuronal diversity: terminal selector genes and selector motifs. Proc. Natl. Acad. Sci. USA 105, 20067–20071.
Hong, J., Hendrix, D.A., Levine, M.S., 2008. Shadow enhancers as a source of evolutionary novelty. Science 321, 1314.
Horan, G.S., Kovàcs, E.N., Behringer, R.R., et al., 1995. Mutations in paralogous HOX genes result in overlapping homeotic transformations of the axial skeleton: evidence for unique and redundant function. Dev. Biol. 169, 359–372.
Hu, J., Hu, H., Li, X., 2008. MOPAT: a graph-based method to predict recurrent cis-regulatory modules from known motifs. Nucleic Acids Res. 36, 4488–4497.
Hufton, A.L., Mathia, S., Braun, H., et al., 2009. Deeply conserved chordate non-coding sequences preserve genome synteny but do not drive gene duplicate retention. Genome Res. 19, 2036–2051.
Irvine, S.Q., Fonseca, V.C., Zompa, M.A., et al., 2008. Cis-regulatory organization of the Pax6 gene in the ascidian Ciona intestinalis. Dev. Biol. 317, 649–659.
Ishihara, T., Sato, S., Ikeda, K., Yajima, H., Kawakami, K., 2008. Multiple evolutionarily conserved enhancers control expression of Eya1. Dev. Dyn. Off. Publ. Am. Assoc. Anatomists 237, 3142–3156.
Ivan, A., Halfon, M.S., Sinha, S., 2008. Computational discovery of cis-regulatory modules in Drosophila without prior knowledge of motifs. Genome Biol. 9, R22.
Iwahori, A., Fraidenraich, D., Basilico, C., 2004. A conserved enhancer element that drives Fgf4 gene expression in the embryonic myotomes is synergistically activated by gata and bhlh proteins. Dev. Biol. 270, 525–537.
Izumi, K., Aramaki, M., Kimura, T., Naito, Y., Udaka, T., Uchikawa, M., Kondoh, H., Suzuki, H., Cho, G., Okada, Y., Takahashi, T., Golden, J.A., Kosaki, K., 2007. Identification of a prosencephalic-specific enhancer of SALL1: comparative genomic approach using the chick embryo. Pediatr. Res. 61, 660–665.
Jack, J., Dorsett, D., Delotto, Y., et al., 1991. Expression of the cut locus in the Drosophila wing margin is required for cell type specification and is regulated by a distant enhancer. Development 113, 735–747.
Janssens, H., Hou, S., Jaeger, J., et al., 2006. Quantitative and predictive model of transcriptional control of the Drosophila melanogaster even skipped gene. Nat. Genet. 38, 1159–1165.
Jiang, Y., Matevossian, A., Huang, H., et al., 2008. Isolation of neuronal chromatin from brain tissue. BMC Neurosci. 9, 42.
Jones, E.A., Flavell, R.A., 2005. Distal enhancer elements transcribe intergenic RNA in the IL-10 family gene cluster. J. Immunol. 175, 7437–7446.
Juan, A.H., Ruddle, F.H., 2003. Enhancer timing of HOX gene expression: deletion of the endogenous HOXC8 early enhancer. Development 130, 4823–4834.
Juven-Gershon, T., Kadonaga, J.T., 2010. Regulation of gene expression via the core promoter and the basal transcriptional machinery. Dev. Biol. 339, 225–229.

Juven-Gershon, T., Cheng, S., Kadonaga, J.T., 2006. Rational design of a super core promoter that enhances gene expression. Nat. Methods 3, 917–922.
Juven-Gershon, T., Hsu, J., Theisen, J.W., Kadonaga, J.T., 2008. The RNA polymerase II core promoter - the gateway to transcription. Curr. Opin. Cell Biol. 20, 253–259.
Kantorovitz, M.R., Kazemian, M., Kinston, S., et al., 2009. Motif-blind, genome-wide discovery of cis-regulatory modules in Drosophila and mouse. Dev. Cell 17, 568–579.
Kantorovitz, M.R., Robinson, G.E., Sinha, S., 2007. A statistical method for alignment-free comparison of regulatory sequences. Bioinformatics (Oxford, England) 23, i249–i255.
Katzman, S., Kern, A.D., Bejerano, G., Fewell, G., Fulton, L., Wilson, R.K., Salama, S.R., Haussler, D., 2007. Human genome ultraconserved elements are ultraselected. Science 317, 915.
Keightley, P.D., Lercher, M.J., Eyre-Walker, A., 2005. Evidence for widespread degradation of gene control regions in hominid genomes. PLoS Biol. 3, e42.
Keightley, P.D., Gaffney, D.J., 2003. Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. Proc. Natl. Acad. Sci. USA 100, 13402–13406.
Kent, W.J., Sugnet, C.W., Furey, T.S., et al., 2002. The human genome browser at UCSC. Genome Res. 12, 996–1006.
Keys, D.N., Lee, B., Di Gregorio, A., et al., 2005. A saturation screen for cis-acting regulatory DNA in the hox genes of ciona intestinalis. Proc. Natl. Acad. Sci. USA 102, 679–683.
Khandekar, M., Suzuki, N., Lewton, J., Yamamoto, M., Engel, J.D., 2004. Multiple, distant Gata2 enhancers specify temporally and tissue-specific patterning in the developing urogenital system. Mol. Cell. Biol. 24, 10263–10276.
Kikuta, H., Fredman, D., Rinkwitz, S., et al., 2007. Retroviral enhancer detection insertions in zebrafish combined with comparative genomics reveal genomic regulatory blocks - a fundamental feature of vertebrate genomes. Genome Biol. 8 (Suppl. 1), S4.
Kim, T.H., Abdullaev, Z.K., Smith, A.D., et al., 2007. Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell 128, 1231–1245.
Kim, T., Hemberg, M., Gray, J.M., et al., 2010. Widespread transcription at neuronal activity-regulated enhancers. Nature 465, 182–187.
Kimura-Yoshida, C., Kitajima, K., Oda-Ishii, I., Tian, E., Suzuki, M., Yamamoto, M., Suzuki, T., Kobayashi, M., Aizawa, S., Matsuo, I., 2004. Characterization of the pufferfish Otx2 cis-regulators reveals evolutionarily conserved genetic mechanisms for vertebrate head specification. Development (Cambridge, England) 131, 57–71.
King, D.C., Taylor, J., Zhang, Y., et al., 2007. Finding cis-regulatory elements using comparative genomics: some lessons from encode data. Genome Res. 17, 775–786.
Kirchhamer, C.V., Yuh, C.H., Davidson, E.H., 1996. Modular cis-regulatory organization of developmentally expressed genes: two genes transcribed territorially in the sea urchin embryo, and additional examples. Proc. Natl Acad. Sci. USA 93, 9322–9328.
Kleinjan, D., Coutinho, P., 2009. Cis-ruption mechanisms: disruption of cis-regulatory control as a cause of human genetic disease. Brief. Funct. Genomics Proteomics 8, 317–332.
Kleinjan, D.A., Seawright, A., Childs, A.J., van Heyningen, V., 2004. Conserved elements in Pax6 intron 7 involved in (auto)regulation and alternative transcription. Dev. Biol. 265, 462–477.
Kobayashi, A., Watanabe, Y., Akasaka, K., et al., 2007. Real-time monitoring of functional interactions between upstream and core promoter sequences in living cells of sea urchin embryos. Nucleic Acids Res. 35, 4882–4894.
Kulkarni, M.M., Arnosti, D.N., 2003. Information display by transcriptional enhancers. Development 130, 6569–6575.
Kural, D., Ding, Y., Wu, J., Korpi, A.M., Chuang, J.H., 2009. COMIT: identification of noncoding motifs under selection in coding sequences. Genome Biol. 10, R133.
Lampe, X., Samad, O.A., Guiguen, A., et al., 2008. An ultraconserved hox-pbx responsive element resides in the coding sequence of hoxa2 and is active in rhombomere 4. Nucleic Acids Res. 36, 3214–3225.
Lee, A.M., Wu, C., 2006. Enhancer–promoter communication at the yellow gene of Drosophila melanogaster: diverse promoters participate in and regulate trans interactions. Genetics 174, 1867–1880.
Lee, A.P., Koh, E.G.L., Tay, A., et al., 2006. Highly conserved syntenic blocks at the vertebrate hox loci and conserved regulatory elements within and outside hox gene clusters. Proc. Natl. Acad. Sci. USA 103, 6994–6999.
Lee, R., 2005. Web resources for c. elegans studies. WormBook 1–16.
Lettice, L.A., Heaney, S.J.H., Purdie, L.A., et al., 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.
Lettice, L.A., Hill, A.E., Devenney, P.S., et al., 2008. Point mutations in a distant sonic hedgehog cis-regulator generate a variable regulatory output responsible for preaxial polydactyly. Hum. Mol. Genet. 17, 978–985.
Li Song, D., Joyner, A.L., 2000. Two Pax2/5/8-binding sites in Engrailed2 are required for proper initiation of endogenous mid-hindbrain expression. Mech. Dev. 90, 155–165.
Li, Q., Barkess, G., Qian, H., 2006. Chromatin looping and the probability of transcription. Trends Genet. 22, 197–202.
Li, X., Noll, M., 1994. Compatibility between enhancers and promoters determines the transcriptional specificity of gooseberry and gooseberry neuro in the Drosophila embryo. EMBO J. 13, 400–406.
Li, X., MacArthur, S., Bourgon, R., et al., 2008. Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. PLoS Biol. 6, e27.
Li, X., Tan, L., Wang, L., et al., 2009. Isolation and characterization of conserved non-coding sequences among rice (Oryza sativa l.) paralogous regions. Mol. Genet. Genomics 281, 11–18.
Li, L., Zhu, Q., He, X., et al., 2007. Large-scale analysis of transcriptional cis-regulatory modules reveals both common features and distinct subclasses. Genome Biol. 8, R101.
Lifanov, A.P., Makeev, V.J., Nazina, A.G., et al., 2003. Homotypic regulatory clusters in Drosophila. Genome Res. 13, 579–588.
Lomvardas, S., Barnea, G., Pisapia, D.J., et al., 2006. Interchromosomal interactions and olfactory receptor choice. Cell 126, 403–413.
Long, Q., Meng, A., Wang, H., et al., 1997. Gata-1 expression pattern can be recapitulated in living transgenic zebrafish using gfp reporter gene. Development 124, 4105–4111.
Long, X., Miano, J.M., 2007. Remote control of gene expression. J. Biol. Chem. 282, 15941–15945.
Loots, G.G., 2008. Genomic identification of regulatory elements by evolutionary sequence comparison and functional analysis. Adv. Genet. 61, 269–293.
Lower, K.M., Hughes, J.R., De Gobbi, M., Henderson, S., Viprakasit, V., Fisher, C., Goriely, A., Ayyub, H., Sloane-Stanley, J., Vernimmen, D., Langford, C., Garrick, D., Gibbons, R.J., Higgs, D.R., 2009. Adventitious changes in long-range gene expression caused by polymorphic structural variation and promoter competition. Proc. Natl Acad. Sci. USA 106, 21771–21776.
MacIsaac, K.D., Fraenkel, E., 2006. Practical strategies for discovering regulatory DNA sequence motifs. PLoS Comput. Biol. 2, e36.
Maconochie, M.K., Nonchev, S., Studer, M., Chan, S.K., Pöpperl, H., Sham, M.H., Mann, R.S., Krumlauf, R., 1997. Cross-regulation in the mouse HoxB complex: the expression of Hoxb2 in rhombomere 4 is regulated by Hoxb1. Genes Dev. 11, 1885–1895.
Maeda, R.K., Karch, F., 2007. Making connections: boundaries and insulators in Drosophila. Curr. Opin. Genet. Dev. 17, 394–399.
Makeev, V.J., Lifanov, A.P., Nazina, A.G., et al., 2003. Distance preferences in the arrangement of binding motifs and hierarchical levels in organization of transcription regulatory information. Nucleic Acids Res. 31, 6016–6026.
Margolin, A.A., Palomero, T., Sumazin, P., Califano, A., Ferrando, A.A., Stolovitzky, G., 2009. ChIP-on-chip significance analysis reveals large-scale binding and regulation by human transcription factor oncogenes. Proc. Natl Acad. Sci. USA 106, 244–249.
Markstein, M., Markstein, P., Markstein, V., et al., 2002. Genome-wide analysis of clustered dorsal binding sites identifies putative target genes in the Drosophila embryo. Proc. Natl. Acad. Sci. USA 99, 763–768.
Markstein, M., Pitsouli, C., Villalta, C., et al., 2008. Exploiting position effects and the gypsy retrovirus insulator to engineer precisely expressed transgenes. Nat. Genet. 40, 476–483.
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., Toublanc, J.E., Petit, C., 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.
Marson, A., Levine, S.S., Cole, M.F., Frampton, G.M., Brambrink, T., Johnstone, S., Guenther, M.G., Johnston, W.K., Wernig, M., Newman, J., Calabrese, J.M., Dennis, L.M., Volkert, T.L., Gupta, S., Love, J., Hannett, N., Sharp, P.A., Bartel, D.P., Jaenisch, R., Young, R.A., 2008. Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. Cell 134, 521–533.
Mahony, S., Benos, P.V., 2007. Stamp: a web tool for exploring DNA-binding motif similarities. Nucleic Acids Res. 35, W253–W258.
Maston, G.A., Evans, S.K., Green, M.R., 2006. Transcriptional regulatory elements in the human genome. Annu. Rev. Genomics Hum. Genet. 7, 29–59.
Matys, V., Fricke, E., Geffers, R., et al., 2003. Transfac: transcriptional regulation, from patterns to profiles. Nucleic Acids Res. 31, 374–378.
McEwen, G.K., Woolfe, A., Goode, D., et al., 2006. Ancient duplicated conserved noncoding elements in vertebrates: a genomic and functional analysis. Genome Res. 16, 451–465.
McGaughey, D.M., Vinton, R.M., Huynh, J., et al., 2008. Metrics of sequence constraint overlook regulatory sequences in an exhaustive analysis at phox2b. Genome Res. 18, 252–260.
McLean, C., Bejerano, G., 2008. Dispensability of mammalian DNA. Genome Res. 18, 1743–1751.
McLellan, A.S., Kealey, T., Langlands, K., 2006. An E box in the exon 1 promoter regulates Insulin-like Growth Factor-I expression in differentiating muscle cells. Am. J. Physiol. Cell Physiol. 291, C300–C307.
Merli, C., Bergstrom, D.E., Cygan, J.A., et al., 1996. Promoter specificity mediates the independent regulation of neighboring genes. Genes Dev. 10, 1260–1270.
Minovitsky, S., Stegmaier, P., Kel, A., Kondrashov, A.S., Dubchak, I., 2007. Short sequence motifs, overrepresented in mammalian conserved non-coding sequences. BMC Genomics 8, 378.
Morachis, J.M., Murawsky, C.M., Emerson, B.M., 2010. Regulation of the p53 transcriptional response by structurally diverse core promoters. Genes Dev. 24, 135–147.
Muffato, M., Louis, A., Poisnel, C.E., Crollius, H.R., 2010. Genomicus: a database and a browser to study gene synteny in modern and ancestral genomes. Bioinformatics 26, 1119–1121.
Navas, P.A., Li, Q., Peterson, K.R., Stamatoyannopoulos, G., 2006. Investigations of a human embryonic globin gene silencing element using YAC transgenic mice. Experimental Biology and Medicine (Maywood, N.J.) 231, 328–334.
Navratilova, P., Fredman, D., Hawkins, T.A., et al., 2009. Systematic human/zebrafish comparative identification of cis-regulatory activity around vertebrate developmental transcription factor genes. Dev. Biol. 327, 526–540.
Nelson, C.E., Hersh, B.M., Carroll, S.B., 2004. The regulatory content of intergenic DNA shapes genome architecture. Genome Biol. 5, R25.
Newburger, D.E., Bulyk, M.L., 2009. UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. Nucleic Acids Res. 37, D77–D82.

Ni, J., Liu, L., Binari, R., Hardy, R., Shim, H., Cavallaro, A., Booker, M., Pfeiffer, B.D., Markstein, M., Wang, H., Villalta, C., Laverty, T.R., Perkins, L.A., Perrimon, N., 2009. A Drosophila resource of transgenic RNAi lines for neurogenetics. Genetics 182, 1089–1100.
Nishihara, H., Smit, A.F.A., Okada, N., 2006. Functional noncoding sequences derived from sines in the mammalian genome. Genome Res. 16, 864–874.
Nóbrega, M.A., Ovcharenko, I., Afzal, V., Rubin, E.M., 2003. Scanning human gene deserts for long-range enhancers. Science 302, 413.
Nóbrega, M.A., Pennacchio, L.A., 2004. Comparative genomic analysis as a tool for biological discovery. J. Physiol. Lond. 554, 31–39.
Nóbrega, M.A., Zhu, Y., Plajzer-Frick, I., et al., 2004. Megabase deletions of gene deserts result in viable mice. Nature 431, 988–993.
Nolis, I.K., McKay, D.J., Mantouvalou, E., Lomvardas, S., Merika, M., Thanos, D., 2009. Transcription factors mediate long-range enhancer–promoter interactions. Proc. Natl Acad. Sci. USA 106, 20222–20227.
Noyes, M.B., Christensen, R.G., Wakabayashi, A., Stormo, G.D., Brodsky, M.H., Wolfe, S.A., 2008. Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. Cell 133, 1277–1289.
Oda-Ishii, I., Bertrand, V., Matsuo, I., Lemaire, P., 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 (Cambridge, England) 132, 1663–1674.
Ohler, U., Wassarman, D.A., 2010. Promoting developmental transcription. Development (Cambridge, England) 137, 15–26.
Ohtsuki, S., Levine, M., Cai, H.N., 1998. Different core promoters possess distinct regulatory activities in the Drosophila embryo. Genes Dev. 12, 547–556.
Oliver, B., Parisi, M., Clark, D., 2002. Gene expression neighborhoods. J. Biol. 1, 4.
Ovcharenko, I., Loots, G.G., Nobrega, M.A., et al., 2005. Evolution and functional classification of vertebrate gene deserts. Genome Res. 15, 137–145.
Palin, K., Taipale, J., Ukkonen, E., 2006. Locating potential enhancer elements by comparative genomics using the EEL software. Nat. Protoc. 1, 368–374.
Panne, D., Maniatis, T., Harrison, S.C., 2007. An atomic model of the interferon-beta enhanceosome. Cell 129, 1111–1123.
Papachatzopoulou, A., Kaimakis, P., Pourfarzad, F., et al., 2007. Increased gamma-globin gene expression in beta-thalassemia intermedia patients correlates with a mutation in 3′hs1. Am. J. Hematol. 82, 1005–1009.
Papatsenko, D., 2007. ClusterDraw web server: a tool to identify and visualize clusters of binding motifs for transcription factors. Bioinformatics (Oxford, England) 23, 1032–1034.
Papatsenko, D., Levine, M., 2005. Computational identification of regulatory DNAs underlying animal development. Nat. Meth. 2, 529–534.
Papatsenko, D., Kislyuk, A., Levine, M., Dubchak, I., 2006. Conservation patterns in different functional sequence categories of divergent Drosophila species. Genomics 88, 431–442.
Papatsenko, D., Goltsev, Y., Levine, M., 2009. Organization of developmental enhancers in the Drosophila embryo. Nucleic Acids Res. 37, 5665–5677.
Park, H.C., Kim, C.H., Bae, Y.K., et al., 2000. Analysis of upstream elements in the huc promoter leads to the establishment of transgenic zebrafish with fluorescent neurons. Dev. Biol. 227, 279–293.
Pena, R.N., Whitelaw, C.B.A., 2005. Duplication of stat5-binding sites within the beta-lactoglobulin promoter compromises transcription in vivo. Biochimie 87, 523–528.
Pennacchio, L.A., Ahituv, N., Moses, A.M., et al., 2006. In vivo enhancer analysis of human conserved non-coding sequences. Nature 444, 499–502.
Phillips, J.E., Corces, V.G., 2009. CTCF: master weaver of the genome. Cell 137, 1194–1211.
Pierstorff, N., Bergman, C.M., Wiehe, T., 2006. Identifying cis-regulatory modules by combining comparative and compositional analysis of DNA. Bioinformatics 22, 2858–2864.
Potts, W., Tucker, D., Wood, H., et al., 2000. Chicken beta-globin 5′HS4 insulators function to reduce variability in transgenic founder mice. Biochem. Biophys. Res. Commun. 273, 1015–1018.
Poulin, F., Nobrega, M.A., Plajzer-Frick, I., et al., 2005. In vivo characterization of a vertebrate ultraconserved enhancer. Genomics 85, 774–781.
Prabhakar, S., Poulin, F., Shoukry, M., et al., 2006. Close sequence comparisons are sufficient to identify human cis-regulatory elements. Genome Res. 16, 855–863.
Putnam, N.H., Butts, T., Ferrier, D.E.K., et al., 2008. The amphioxus genome and the evolution of the chordate karyotype. Nature 453, 1064–1071.
Rabinovich, A., Jin, V.X., Rabinovich, R., et al., 2008. E2F in vivo binding specificity: comparison of consensus versus nonconsensus binding sites. Genome Res. 18, 1763–1777.
Rahimov, F., Marazita, M.L., Visel, A., et al., 2008. Disruption of an AP-2alpha binding site in an Irf6 enhancer is associated with cleft lip. Nat. Genet. 40, 1341–1347.
Rajewsky, N., Vergassola, M., Gaul, U., Siggia, E.D., 2002. Computational detection of genomic cis-regulatory modules applied to body patterning in the early Drosophila embryo. BMC Bioinform. 3, 30.
Rebeiz, M., Reeves, N.L., Posakony, J.W., 2002. Score: a computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data. site clustering over random expectation. Proc. Natl Acad. Sci. USA 99, 9888–9893.
Retelska, D., Beaudoing, E., Notredame, C., et al., 2007. Vertebrate conserved non coding DNA regions have a high persistence length and a short persistence time. BMC Genomics 8, 398.
Rinn, J.L., Kertesz, M., Wang, J.K., et al., 2007. Functional demarcation of active and silent chromatin domains in human hox loci by noncoding RNAs. Cell 129, 1311–1323.
Roider, H.G., Manke, T., O'Keeffe, S., Vingron, M., Haas, S.A., 2009. PASTAA: identifying transcription factors associated with sets of co-regulated genes. Bioinformatics (Oxford, England) 25, 435–442.
Roider, H.G., Kanhere, A., Manke, T., et al., 2007. Predicting transcription factor affinities to DNA from a biophysical model. Bioinformatics 23, 134–141.
Romano, L.A., 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.
Ronshaugen, M., Levine, M., 2004. Visualization of trans-homolog enhancer–promoter interactions at the abd-b hox locus in the Drosophila embryo. Dev. Cell 7, 925–932.
Sabherwal, N., Bangs, F., Röth, R., et al., 2007. Long-range conserved non-coding SHOX sequences regulate expression in developing chicken limb and are associated with short stature phenotypes in human patients. Hum. Mol. Genet. 16, 210–222.
Sakuraba, Y., Kimura, T., Masuya, H., et al., 2008. Identification and characterization of new long conserved noncoding sequences in vertebrates. Mamm. Genome 19, 703–712.
Sanchez-Elsner, T., Gou, D., Kremmer, E., et al., 2006. Noncoding RNAs of trithorax response elements recruit Drosophila ash1 to ultrabithorax. Science 311, 1118–1123.
Sandelin, A., Bailey, P., Bruce, S., et al., 2004. Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. BMC Genomics 5, 99.
Sandve, G.K., Drabløs, F., 2006. A survey of motif discovery methods in an integrated framework. Biol. Direct 1, 11.
Santagati, F., Abe, K., Schmidt, V., et al., 2003. Identification of cis-regulatory elements in the mouse Pax9/nkx2-9 genomic region: implication for evolutionary conserved synteny. Genetics 165, 235–242.
De Santa, F., Barozzi, I., Mietton, F., et al., 2010. A large fraction of extragenic RNA pol II transcription sites overlap enhancers. PLoS Biol. 8, e1000384.
Satija, R., Pachter, L., Hein, J., 2008. Combining statistical alignment and phylogenetic footprinting to detect regulatory elements. Bioinformatics 24, 1236–1242.
Schlabach, M.R., Hu, J.K., Li, M., et al., 2010. Synthetic design of strong promoters. Proc. Natl. Acad. Sci. USA 107, 2538–2543.
Schroeder, M.D., Pearce, M., Fak, J., et al., 2004. Transcriptional control in the segmentation gene network of Drosophila. PLoS Biol. 2, E271.
Segal, E., Raveh-Sadka, T., Schroeder, M., et al., 2008. Predicting expression patterns from regulatory sequence in Drosophila segmentation. Nature 451, 535–540.
Sharan, S.K., Thomason, L.C., Kuznetsov, S.G., et al., 2009. Recombineering: a homologous recombination-based method of genetic engineering. Nat. Protoc. 4, 206–223.
Shin, J.T., Priest, J.R., Ovcharenko, I., Ronco, A., Moore, R.K., Burns, C.G., 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., et al., 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15, 1034–1050.
Sierro, N., Kusakabe, T., Park, K., et al., 2006. Dbtgr: a database of tunicate promoters and their regulatory elements. Nucleic Acids Res. 34, D552–D555.
Simonis, M., Klous, P., Splinter, E., et al., 2006. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4c). Nat. Genet. 38, 1348–1354.
Sinha, S., Siggia, E.D., 2005. Sequence turnover and tandem repeats in cis-regulatory modules in Drosophila. Mol. Biol. Evol. 22, 874–885.
Smale, S.T., 2001. Core promoters: active contributors to combinatorial gene regulation. Genes Dev. 15, 2503–2508.
Smith, A.D., Sumazin, P., Zhang, M.Q., 2007. Tissue-specific regulatory elements in mammalian promoters. Mol. Syst. Biol. 3, 73.
Smith, J., 2008. A protocol describing the principles of cis-regulatory analysis in the sea urchin. Nat. Protoc. 3, 710–718.
Soshnikova, N., Duboule, D., 2009. Epigenetic temporal control of mouse Hox genes in vivo. Science 324, 1320–1323.
Spitz, F., Duboule, D., 2008. Global control regions and regulatory landscapes in vertebrate development and evolution. Adv. Genet. 61, 175–205.
Spitz, F., Gonzalez, F., Duboule, D., 2003. A global control region defines a chromosomal regulatory landscape containing the HoxD cluster. Cell 113, 405–417.
Spradling, A.C., Rubin, G.M., 1983. The effect of chromosomal position on the expression of the Drosophila xanthine dehydrogenase gene. Cell 34, 47–57.
Stephen, S., Pheasant, M., Makunin, I.V., et al., 2008. Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock. Mol. Biol. Evol. 25, 402–408.
Sun, H., Skogerbø, G., Chen, R., 2006. Conserved distances between vertebrate highly conserved elements. Hum. Mol. Genet. 15, 2911–2922.
Suster, M.L., Kania, A., Liao, M., et al., 2009. A novel conserved Evx1 enhancer links spinal interneuron morphology and cis-regulation from fish to mammals. Dev. Biol. 325, 422–433.
Suzuki, Y., Yamashita, R., Nakai, K., et al., 2002. DBTSS: database of human transcriptional start sites and full-length cDNAs. Nucleic Acids Res. 30, 328–331.
Taher, L., Ovcharenko, I., 2009. Variable locus length in the human genome leads to ascertainment bias in functional inference for non-coding elements. Bioinformatics 25, 578–584.
Tassy, O., Daian, F., Hudson, C., et al., 2006. A quantitative approach to the study of cell shapes and interactions during early chordate embryogenesis. Curr. Biol. 16, 345–358.
Taylor, J., Tyekucheva, S., King, D.C., et al., 2006. ESPERR: learning strong and weak signals in genomic sequence alignments to identify functional elements. Genome Res. 16, 1596–1604.
Thanos, D., Maniatis, T., 1995. Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome. Cell 83, 1091–1100.

Tompa, M., Li, N., Bailey, T.L., et al., 2005. Assessing computational tools for the discovery of transcription factor binding sites. Nat. Biotechnol. 23, 137–144.
Tronche, F., Ringeisen, F., Blumenfeld, M., et al., 1997. Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome. J. Mol. Biol. 266, 231–245.
Tsang, W.H., Shek, K.F., Lee, T.Y., et al., 2009. An evolutionarily conserved nested gene pair- mab21 and lrba/nbea in metazoan. Genomics.
Tümpel, S., Cambronero, F., Sims, C., et al., 2008. A regulatory module embedded in the coding region of hoxa2 controls expression in rhombomere 2. Proc. Natl. Acad. Sci. USA 105, 20077–20082.
Tursun, B., Cochella, L., Carrera, I., et al., 2009. A toolkit and robust pipeline for the generation of fosmid-based reporter genes in c. elegans. PLoS ONE 4, e4625.
Uchikawa, M., Ishida, Y., Takemoto, T., et al., 2003. Functional analysis of chicken sox2 enhancers highlights an array of diverse regulatory elements that are conserved in mammals. Dev. Cell 4, 509–519.
Vakoc, C.R., Letting, D.L., Gheldof, N., et al., 2005. Proximity among distant regulatory elements at the beta-globin locus requires gata-1 and fog-1. Mol. Cell 17, 453–462.
Vandenbon, A., Nakai, K., 2010. Modeling tissue-specific structural patterns in human and mouse promoters. Nucleic Acids Res. 38, 17–25.
Vavouri, T., Elgar, G., 2005. Prediction of cis-regulatory elements using binding site matrices–the successes, the failures and the reasons for both. Curr. Opin. Genet. Dev. 15, 395–402.
Vavouri, T., McEwen, G.K., Woolfe, A., et al., 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., et al., 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.
Venken, K.J.T., Carlson, J.W., Schulze, K.L., et al., 2009. Versatile p[acman] bac libraries for transgenesis studies in drosophila melanogaster. Nat. Methods 6, 431–434.
Visel, A., Akiyama, J.A., Shoukry, M., et al., 2009a. Functional autonomy of distant-acting human enhancers. Genomics 93, 509–513.
Visel, A., Blow, M.J., Li, Z., et al., 2009b. ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature 457, 854–858.
Visel, A., Bristow, J., Pennacchio, L.A., 2007. Enhancer identification through comparative genomics. Semin. Cell Dev. Biol. 18, 140–152.
Voth, H., Oberthuer, A., Simon, T., et al., 2009. Co-regulated expression of Hand2 and Dein by a bidirectional promoter with asymmetrical activity in neuroblastoma. BMC Mol. Biol. 10, 28.
Wakaguri, H., Yamashita, R., Suzuki, Y., et al., 2008. DBTSS: database of transcription start sites, progress report 2008. Nucleic Acids Res. 36, D97–D101.
Walter, K., Abnizova, I., Elgar, G., et al., 2005. Striking nucleotide frequency pattern at the borders of highly conserved vertebrate non-coding sequences. Trends Genet. 21, 436–440.
Wang, H., Zhang, Y., Cheng, Y., et al., 2006. Experimental validation of predicted mammalian erythroid cis-regulatory modules. Genome Res. 16, 1480–1492.
Wang, Q., Prabhakar, S., Chanan, S., et al., 2007a. Detection of weakly conserved ancestral mammalian regulatory sequences by primate comparisons. Genome Biol. 8, R1.
Wang, T., Chen, Y., Liu, C., et al., 2002. Functional analysis of the proximal promoter regions of fish Rhodopsin and MYF-5 genes using transgenesis. Mar. Biotechnol. 4, 247–255.
Wang, W., Zhong, J., Su, B., et al., 2007b. Comparison of Pax1/9 locus reveals 500-myr-old syntenic block and evolutionary conserved noncoding regions. Mol. Biol. Evol. 24, 784–791.
Washietl, S., Pedersen, J.S., Korbel, J.O., et al., 2007. Structured RNAs in the ENCODE selected regions of the human genome. Genome Res. 17, 852–864.
Wasserman, W.W., Fickett, J.W., 1998. Identification of regulatory regions which confer muscle-specific gene expression. J. Mol. Biol. 278, 167–181.
Wasserman, W.W., Sandelin, A., 2004. Applied bioinformatics for the identification of regulatory elements. Nat. Rev. Genet. 5, 276–287.
Waterston, R.H., Lindblad-Toh, K., Birney, E., et al., 2002. Initial sequencing and comparative analysis of the mouse genome. Nature 420, 520–562.
Wederell, E.D., Bilenky, M., Cullum, R., et al., 2008. Global analysis of in vivo foxa2-binding sites in mouse adult liver using massively parallel sequencing. Nucleic Acids Res. 36, 4549–4564.
Wenick, A.S., Hobert, O., 2004. Genomic cis-regulatory architecture and trans-acting regulators of a single interneuron-specific gene battery in c. elegans. Dev. Cell 6, 757–770.
Werner, T., Hammer, A., Wahlbuhl, M., Bösl, M.R., Wegner, M., 2007. Multiple conserved regulatory elements with overlapping functions determine Sox10 expression in mouse embryogenesis. Nucleic Acids Res. 35, 6526–6538.
West, A.G., Fraser, P., 2005. Remote control of gene transcription. Hum. Mol. Genet. 14 (Spec. No. 1), R101–R111.
Woltering, J.M., Duboule, D., 2009. Conserved elements within open reading frames of mammalian HOX genes. J. Biol. 8, 17.
Woolfe, A., Elgar, G., 2007. Comparative genomics using fugu reveals insights into regulatory subfunctionalization. Genome Biol. 8, R53.
Woolfe, A., Elgar, G., 2008. Organization of conserved elements near key developmental regulators in vertebrate genomes. Adv. Genet. 61, 307–338.
Woolfe, A., Goodson, M., Goode, D.K., et al., 2005. Highly conserved non-coding sequences are associated with vertebrate development. PLoS Biol. 3, e7.
Wratten, N.S., McGregor, A.P., Shaw, P.J., et al., 2006. Evolutionary and functional analysis of the tailless enhancer in musca domestica and Drosophila melanogaster. Evol. Dev. 8, 6–15.
Xie, X., Kamal, M., Lander, E.S., 2006. A family of conserved noncoding elements derived from an ancient transposable element. Proc. Natl. Acad. Sci. USA 103, 11659–11664.
Xiong, L., Catoire, H., Dion, P., et al., 2009. Meis1 intronic risk haplotype associated with restless legs syndrome affects its mRNA and protein expression levels. Hum. Mol. Genet. 18, 1065–1074.
Xiong, N., Kang, C., Raulet, D.H., 2002. Redundant and unique roles of two enhancer elements in the Tcrgamma locus in gene regulation and Gammadelta T cell development. Immunity 16, 453–463.
Xu, X., Bieda, M., Jin, V.X., Rabinovich, A., Oberley, M.J., Green, R., Farnham, P.J., 2007. A comprehensive ChIP-chip analysis of E2F1, E2F4, and E2F6 in normal and tumor cells reveals interchangeable roles of E2F family members. Genome Res. 17, 1550–1561.
Xu, X., Scott, M.M., Deneris, E.S., 2006. Shared long-range regulatory elements coordinate expression of a gene cluster encoding nicotinic receptor heteromeric subtypes. Mol. Cell. Biol. 26, 5636–5649.
Yanagisawa, H., Clouthier, D.E., Richardson, J.A., et al., 2003. Targeted deletion of a branchial arch-specific enhancer reveals a role of Dhand in craniofacial development. Development 130, 1069–1078.
Yang, G.S., Banks, K.G., Bonaguro, R.J., et al., 2009. Next generation tools for high-throughput promoter and expression analysis employing single-copy knock-ins at the HPRT1 locus. Genomics 93, 196–204.
Yoshikawa, S., Norcom, E., Nakamura, H., et al., 2007. Transgenic analysis of the anterior eye-specific enhancers of the zebrafish Gelsolin-like 1 (gsnl1) gene. Dev. Dyn. 236, 1929–1938.
Yuan, Y., Guo, L., Shen, L., et al., 2007. Predicting gene expression from sequence: a reexamination. PLoS Comput. Biol. 3, e243.
Yuh, C.H., Bolouri, H., Davidson, E.H., 1998. Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. Science 279, 1896–1902.
Zhang, C., Xuan, Z., Otto, S., et al., 2006. A clustering property of highly-degenerate transcription factor binding sites in the mammalian genome. Nucleic Acids Res. 34, 2238–2246.
Zhang, Z.D., Paccanaro, A., Fu, Y., et al., 2007. Statistical analysis of the genomic distribution and correlation of regulatory elements in the encode regions. Genome Res. 17, 787–797.
Zheng, J.B., Zhou, Y.H., Maity, T., et al., 2001. Activation of the human PAX6 gene through the exon 1 enhancer by transcription factors SEF and Sp1. Nucleic Acids Res. 29, 4070–4078.
Zinzen, R.P., Girardot, C., Gagneur, J., Braun, M., Furlong, E.E.M., 2009. Combinatorial binding predicts spatio-temporal cis-regulatory activity. Nature 462, 65–70 [NCICRF]. http://recombineering.ncifcrf.gov.