

# Genomic Identification of Regulatory Elements by Evolutionary Sequence Comparison and Functional Analysis

---

**Gabriela G. Loots**

Biosciences and Biotechnology Division, Chemistry, Materials and Life Sciences Directorate, Lawrence Livermore National Laboratory, Livermore, California 94550

- I. Introduction
  - II. Genomic Architecture of the Human Genome
    - A. Distant regulatory elements controlling transcription
    - B. Noncoding mutations causing human disease
  - III. Computational Methods of Predicting Regulatory Elements
    - A. Identifying evolutionarily conserved noncoding sequences
    - B. Predicting TFBSs
  - IV. *In Vivo* Validation and Characterization of Transcriptional Regulatory Elements
    - A. Enhancer validation using transient transgenesis
    - B. Identifying distant enhancers using large genomic constructs
    - C. Mutating candidate regulatory elements in engineered mice
  - V. Conclusions
- References

---

## ABSTRACT

Despite remarkable recent advances in genomics that have enabled us to identify most of the genes in the human genome, comparable efforts to define transcriptional *cis*-regulatory elements that control gene expression are lagging behind. The difficulty of this task stems from two equally important problems: our knowledge of how regulatory elements are encoded in genomes remains

elementary, and there is a vast genomic search space for regulatory elements, since most of mammalian genomes are noncoding. Comparative genomic approaches are having a remarkable impact on the study of transcriptional regulation in eukaryotes and currently represent the most efficient and reliable methods of predicting noncoding sequences likely to control the patterns of gene expression. By subjecting eukaryotic genomic sequences to computational comparisons and subsequent experimentation, we are inching our way toward a more comprehensive catalog of common regulatory motifs that lie behind fundamental biological processes. We are still far from comprehending how the transcriptional regulatory code is encrypted in the human genome and providing an initial global view of regulatory gene networks, but collectively, the continued development of comparative and experimental approaches will rapidly expand our knowledge of the transcriptional regulome. © 2008, Elsevier Inc.

---

## I. INTRODUCTION

In contrast to the genomic landscape of many prokaryotic organisms that are compact and gene rich, most eukaryotic, particularly metazoan, genomes have a small ratio of genes to noncoding DNA since only a minority of the genome is transcribed and translated into proteins. The focus of the initial analysis of both the human (Lander *et al.*, 2001; Venter *et al.*, 2001) and mouse genomes (Waterston *et al.*, 2002) has been to catalog all mammalian protein-coding genes, now estimated to be in the vicinity of  $\sim 25,000$  unique transcripts (Collins, 2004), and spanning less than 2% of the human genome. An additional 40–45% of the human genome is covered by repetitive DNA elements, while the remaining  $\sim 53\%$  is composed of noncoding DNA (Lander *et al.*, 2001; Venter *et al.*, 2001; Waterston *et al.*, 2002). Despite this vast amount of noncoding DNA, little progress has been made in conclusively determining whether it plays any vital functional role. Although some parts of noncoding regions within our genome will eventually reveal no detectable biological function, a growing hypothesis speculates that much of an organism's genetic complexity is due to elaborate transcriptional regulatory signals embedded in our noncoding DNA that determine when, where, and what amount of a gene transcript is expressed.

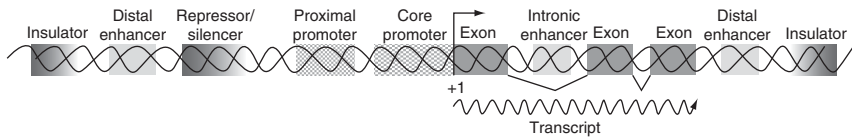
Natural selection is a major driving force in stabilizing functionally important regions within genomes, preserving the sequences of orthologous coding exons and transcriptional regulatory elements. Evolutionary comparisons have long held the promise for identifying transcriptional response elements in eukaryotic genomes, where initially searches were conducted using consensus sequences and positional weight matrices in a method often termed *phylogenetic footprinting* (Blanchette *et al.*, 2002; Chiang *et al.*, 2003; Fink *et al.*, 1996; Wasserman and Fickett, 1998). A new generation of *ab initio* approaches have

increasingly shown great potential for identifying novel functional motifs, but in general the sheer size and complexity of mammalian genomes has precluded the extension of these approaches to study mammalian transcription on a genome scale. This has been primarily because these simple motifs are short and highly degenerate and create overwhelming predictions with high rates of false positives when used in whole-genome analysis. The availability of large amounts of sequence data from numerous organisms and new user-friendly alignment tools have totally altered our contemporary approach to transcriptional regulation, where evolutionary comparisons have become the first tier of analysis routinely performed when searching for regulatory elements. This is reflected by the dramatic increase in the number of studies reporting the identification of functional sequences through the use of comparative genomics, and these emerging studies are providing compelling evidence in support for the use of evolutionary comparisons as a robust strategy for highlighting functional coding (Gilligan *et al.*, 2002; Pennacchio *et al.*, 2001) and noncoding sequences (Gottgens *et al.*, 2000; Loots *et al.*, 2000; Nobrega *et al.*, 2003; Pennacchio *et al.*, 2006; Touchman *et al.*, 2000). In particular, aligning whole genomes and identifying evolutionarily conserved regions (ECR) on a large scale has become a robust approach for discovering transcriptional regulatory elements in noncoding DNA (Pennacchio *et al.*, 2006; Woolfe *et al.*, 2005). Here we will discuss methods of applying comparative genomics to the identification of transcriptional regulatory elements in the human genome, and functional approaches for validating and characterizing computationally predicted elements.

---

## II. GENOMIC ARCHITECTURE OF THE HUMAN GENOME

It is not fully understood how one could precisely define a human gene locus, since all functional elements have yet to be determined for each transcript, but in general, one could view a typical animal gene as a promoter linked to the transcript, both of which are embedded in a sea of positively and negatively regulating elements positioned anywhere in relation to the transcriptional start site (5', 3', and intronic), and acting at a distance across large segments of DNA (up to megabases in lengths) (Fig. 10.1). The positively regulating elements or *enhancers* are each responsible for a subset of the total gene expression pattern and usually drive transcription within a specific tissue or subset of cell types. A typical enhancer can range in size from as little as 100 base pairs (bp) (Banet *et al.*, 2000; Catena *et al.*, 2004; Krebsbach *et al.*, 1996) to several kilobases (kb) in length (Chi *et al.*, 2005; Danielian *et al.*, 1997), but on average would be about 500 bp in length (Kamat *et al.*, 1999; Loots *et al.*, 2005; Marshall *et al.*, 2001). Within enhancer elements are docking sites for regulatory proteins or transcription factors (TFs) that physically interact with specific DNA sequences or



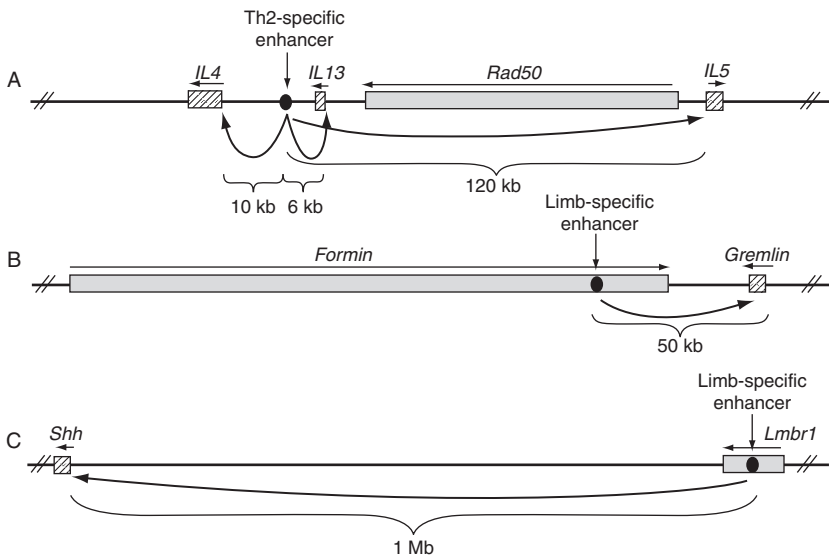
**Figure 10.1.** Schematic representation of a transcriptional locus in the human genome that consists of a complex arrangement of multiple clustered regulatory modules that may include elements such as enhancers, silencers, and insulators which interact with proximal and core promoter elements to specify transcription. These modules can be located across large stretches of DNA up to megabases in length and can be located either upstream or downstream of the transcriptional start site.

transcription factor binding sites (TFBSs). TFs recognize and bind to short (6–12 bp), highly degenerate sequence motifs that occur very frequently in a genome; therefore, computationally predicting TFBSs that are functionally significant is a great challenge. It is not yet known how many TFBSs are needed to build a functional enhancer, nor how many different TFs need to synergistically cooperate to drive expression. One hypothesis suggests that a typical enhancer contains a minimum of 10 TFBSs for at least 3 different TFs (Levine and Tjian, 2003).

The core promoter serves as a platform for the assembly of transcriptional preinitiation complex (PIC) that includes TFIIA, TFIIB, TFIID, TFIIE, TFIIF, TFIIH, and RNA polymerase II (Pol II), which function collectively to specify the transcription start site. The PIC usually begins with TFIID binding to the TATA box (a TATA box is a DNA sequence found in the promoter region of eukaryotic genes, specified as 5'-TATAA-4-3' or a variant), initiator, and/or downstream promoter element (DPE) found in most core promoters, followed by the entry of other general transcription factors (GTFs) and Pol II through either a sequential assembly or a preassembled Pol II holoenzyme pathway (Thomas and Chiang, 2006). This promoter-bound complex is sufficient to drive basal level of transcription, but would require additional cofactors to transmit regulatory signals between gene-specific activators and the general transcription machinery. Three classes of general cofactors, including TATA-binding protein (TBP)-associated factors (TAFs), mediator, and upstream stimulatory activity (USA)-derived positive cofactors (PC1/PARP-1, PC2, PC3/DNA topoisomerase I, and PC4) and negative cofactor 1 (NC1/HMGB1), normally function independently or in combination to fine-tune the promoter activity in a gene-specific or cell-type-specific manner. In addition, other cofactors, such as TAF1, BTAF1, and NC2, can also modulate TBP or TFIID binding to the core promoter. Many genes also contain binding TFBSs for proximal regulatory factors located just 5' of the core promoter. These factors do not always function as classic activators or repressors; instead, they might serve as a connector between distal enhancers and the core promoter. In addition, a different class of regulatory elements, *insulators* or *boundary elements*, serve as gatekeepers and prevent enhancers from inappropriately regulating neighboring genes.

## A. Distant regulatory elements controlling transcription

What makes transcriptional genomics in vertebrates a highly intricate problem stems from two recent observations: (1) all regulatory elements associated with a transcript can be scattered over great distances that can reach megabases (Mb) in length (Nobrega *et al.*, 2003; Sagai *et al.*, 2005), and (2) some regulatory elements are capable of controlling multiple transcripts, skip intercalating genes, or regulate one transcript while being positioned within a different transcript (Loots *et al.*, 2000; Zuniga *et al.*, 2004). In Fig. 10.2 we depict three such examples. For the human interleukin 4 (*IL4*) gene cluster on human chromosome 5, it was long hypothesized that a common regulatory element or locus control region possibly controls the Th2 expression of several cytokine genes. Using comparative genomics one highly conserved element positioned between *IL4* and *IL13* was removed from a human yeast artificial chromosome transgene (Loots *et al.*, 2000) as well as from the mouse genome (Mohrs *et al.*, 2001) to show that by removing this element the expression of three cytokines *IL4*, *IL13* and *IL5* is affected in Th2 cells. What was peculiar about this discovery was the fact that this regulatory element was positioned 120 kb away from the promoter of *IL5* gene, and it was able to exclusively control these three cytokines at the transcriptional level, leaving the intercalating gene, *RAD50*, unaffected when removed from the genome (Fig. 10.2A) (Loots *et al.*, 2000; Mohrs *et al.*, 2001).



**Figure 10.2.** Long-range enhancer activity observed for the interleukin 4, 5, and 13 locus (A), the *formin-gremlin* locus (B), and the *sonic hedgehog-Lmbr1* locus (C).

In the case of limb deformity mutations originally mapped to the C-terminal region of the *formin* gene almost two decades ago (Mass *et al.*, 1990; Woychik and Alagramam, 1998; Woychik *et al.*, 1985, 1990), it was long hypothesized that *formin* is the gene responsible for the disruption of limb bud morphogenesis. Recently when a null mutation in the neighboring gene *gremlin* was generated, it was shown that by removing this BMP antagonist one recapitulates the limb phenotypes recorded for all limb mutations mapped to the *formin* locus (Khokha *et al.*, 2003). Consequent complementation studies together with *in vivo* enhancer expression assays confirmed that the limb deformity phenotypes were indeed a result of *gremlin*-specific regulatory element mutations, and that this limb-specific enhancer resides in an intron, in the C terminus region of the large *formin* transcript, and that *formin* does not contribute to the limb morphogenic abnormalities due to these mutations (Fig. 10.2B) (Zuniga *et al.*, 2004). In a more dramatic example, a region in the fifth intron of *Lmbr1* gene has long been suggested as the responsible element for preaxial polydactyly recorded in mouse and human mutations (Heutink *et al.*, 1994). In these mice, *Shh* expression is perturbed in the anterior margin of the limb bud mesenchyme (Masuya *et al.*, 1995), a gene positioned 1 Mb away from the *Lmbr1* transcript, recapitulating phenotypic aspects of the *Shh* knockout. These studies were followed by transgenic reporter experiments which revealed that this intronic region from *Lmbr1* is indeed an enhancer that drives expression in the posterior mesenchyme of the developing mouse limb bud (Lettice *et al.*, 2003). Most recently, this enhancer was removed from the mouse genome to conclusively show that this element located 1 Mb away from the *Shh* transcriptional start site is required for the limb-specific expression of *Shh*, and its ablation results in a limb phenotype identical to the limb phenotype observed in *Shh* knockout, and therefore it functions as a limb-specific *Shh* enhancer (Fig. 10.2C) (Sagai *et al.*, 2005).

## B. Noncoding mutations causing human disease

In the absence of sequence information, one method that was employed to genetically map congenital abnormalities was to karyotype affected individuals and determine whether chromosomal abnormalities in the form of deletions and translocations segregate with the phenotype in affected families. While generally, detailed mapping and targeted sequencing of affected individuals lead to the discovery of the causative gene and identification of the deleterious mutations, in some instances these chromosomal aberrations do not disrupt any genes or coding regions—several such examples are listed in Table 10.1 (Ahituv *et al.*, 2004). For example, mutations in the coding region of the gene encoding for the developmental transcription factor *SALL1* lead to autosomal dominant Townes–Brocks syndrome, while a thoroughly characterized translocation in one patient 180 kb telomeric to *SALL1* also leads to a similar phenotype

**Table 10.1.** Human Abnormalities Mapped to Noncoding Regions

| Gene        | Disease  | References                          |
|-------------|--|-------------------------------------|
| FOXC2       | Lymphedema-distichiasis                              | (Fang <i>et al.</i> , 2000)         |
| FOXL2       | Blepharophimosis/ptosis/epicanthus inversus syndrome | (Crisponi <i>et al.</i> , 2004)     |
| FSHD        | Facioscapulohumeral dystrophy                        | (van Deutekom <i>et al.</i> , 1996) |
| GLI3        | Greig cephalopolysyndactyly                          | (Vortkamp <i>et al.</i> , 1991)     |
| HBB complex | Gb-Thalassemia                                       | (Kioussis <i>et al.</i> , 1983)     |
| PAX6        | Aniridia   | (Fantes <i>et al.</i> , 1995)       |
| PITX2       | Rieger syndrome                                      | (Flomen <i>et al.</i> , 1997)       |
| PLP1        | Pelizaeus–Merzbacher disease                         | (Inoue <i>et al.</i> , 2002)        |
| POU3F4      | X-linked deafness                                    | (de Kok <i>et al.</i> , 1995)       |
| SALL1       | Townes–Brocks syndrome                               | (Marlin <i>et al.</i> , 1999)       |
| Shh         | Holoprosencephaly                                    | (Roessler <i>et al.</i> , 1997)     |
| SIX3        | Holoprosencephaly                                    | (Wallis <i>et al.</i> , 1999)       |
| SOST        | Van Buchem disease                                   | (Balemans <i>et al.</i> , 2002)     |
| SOX9        | Campomelic displasia                                 | (Wirth <i>et al.</i> , 1996)        |
| SRY         | Sex reversal TWIST Saethre–Chotzen syndrome          | (McElreavy <i>et al.</i> , 1992)    |

(Marlin *et al.*, 1999). One likely explanation for these observations is that noncoding *cis*-regulatory sequences have been mutated or removed from the genome, affecting the expression pattern or expression level of the gene they normally regulate. Since many recorded diseases have no documented coding mutations, it is likely that disruption in the communication between a vital *cis*-regulatory sequence and the gene it regulates could potentially result in a disease that resembles hypomorphic or null alleles of the causative gene. However, providing definitive proof that the noncoding sequence change is indeed causing a particular phenotype is a highly complex problem, difficult to address experimentally. Recently, it has been suggested that engineered bacterial artificial chromosomes (BAC) may be used to determine if noncoding deletions deleteriously impact gene expression of disease-causing genes (Loots *et al.*, 2005). The authors investigated whether a homozygous 52-kb noncoding deletion linked to the sclerosteosis disease-causing gene, *SOST* (Balemans *et al.*, 2001), and homozygous in Van Buchem (VB) patients affects *SOST* gene expression by expressing a wild-type BAC and a genetically modified BAC mimicking the VB allele. They proceeded to show that a *SOST* wild-type allele expresses human *SOST* according to its endogenous expression pattern, primarily in the adult bone, while the VB allele fails to drive *SOST* expression in the bone. They further proceeded to use comparative sequence analysis and enhancer assays to identify a distant enhancer element that is able to drive transgenic expression in osteocyte-like cell lines, and in the mouse skeletal anlage at E14.5 (Loots *et al.*, 2005).

### III. COMPUTATIONAL METHODS OF PREDICTING REGULATORY ELEMENTS

The majority of available computational tools for predicting regulatory elements are based on constructing alignments between orthologous sequences and/or detecting TF DNA binding motifs. Investigators now have the option to deduce phylogenetic relationships among sequences either by generating their own alignments (Bray *et al.*, 2003; Brudno *et al.*, 2003; Mayor *et al.*, 2000; Schwartz *et al.*, 2003) or by using ready-made DNA conservation plots available at various genome browsers (Kuhn *et al.*, 2007; Ovcharenko *et al.*, 2004b; Schwartz *et al.*, 2003). There are several different approaches for scanning sequences for putative regulatory elements using pattern recognition. First, a number of computational tools predict TFBSs using a library of known motifs (Heinemeyer *et al.*, 1998, 1999; Loots and Ovcharenko, 2004), or identify conserved sequence blocks in a multiple sequence alignment (Blanchette *et al.*, 2002; Hertz and Stormo, 1999). Clustering of TFBS has been implemented as a second approach for predicting regulatory elements or *cis*-regulatory modules (CRMs). A few programs analyze homogenous clusters of a single overrepresented DNA motif or heterogenous clusters of multiple different sequence motifs (synergistic motifs) (Berman *et al.*, 2002; Kim *et al.*, 2006; Loots *et al.*, 2002). A third approach for predicting sequences with specific regulatory properties is to identify CRMs shared by multiple functionally related sequences from the same organism (Jegga *et al.*, 2005, 2007; Sharan *et al.*, 2004). Expression profiling experiments have the potential to identify groups of coexpressed genes that respond to similar environmental and metabolic stimuli, and it has been speculated that such gene sets often share similar types of CRMs because their coregulation is mediated by the same set of regulatory proteins. Several new computational approaches use microarray expression data to predict tissue-specific regulatory elements in coexpressed set of genes (Jegga *et al.*, 2005; Ovcharenko and Nobrega, 2005; Pennacchio *et al.*, 2007).

#### A. Identifying evolutionarily conserved noncoding sequences

Since sequences that mediate gene expression tend to be evolutionarily conserved, one can identify putative enhancers by comparing genomes and determining regions of high homology (Loots *et al.*, 2000, 2005; Nobrega *et al.*, 2003). To identify evolutionarily conserved noncoding sequences (ncECRs), one needs to be able to generate reliable alignments between orthologous noncoding regions from different organisms. Aligning short sequences from closely related organisms is a straightforward process, while determining sequence similarity between larger, highly divergent regions is a more difficult task due to significant DNA rearrangements. Even highly orthologous regions are rich in insertions and deletions as well as many single base-pair mutations which can lack orthologous

counterparts and be represented as gaps within alignments. An additional potential difficulty in obtaining accurate syntenic alignments is created by the large numbers of tandem and segmental duplications found in the human genome. Some assembly strategies are unable to differentiate highly homologous duplications from true overlapping sequences, resulting in erroneous genomic assemblies with underrepresented paralogs. The most complex problem is posed by lineage-specific segmental duplications that arose since the separation of two species from their most recent common ancestor. In this situation identifying true orthologous syntenic sequences from paralogous ones is a difficult task since determining true orthology and synteny represents a major challenge in the absence of a one-to-one sequence match.

The majority of pairwise sequence alignment programs utilize dynamic programming of global alignments (Needleman and Wunsch, 1970), local alignments (Altschul *et al.*, 1990), or database searches (Altschul *et al.*, 1997). A small fraction of alignment programs use hidden Markov models (HMM) such as WABA (Kent and Zahler, 2000) or suffix tree such as MUMmer (Majoros *et al.*, 2005) and AVID (Bray *et al.*, 2003). In a very simplistic view, global alignments assume that there is a colinearity of DNA sequences while local alignments focus on detecting short matches between two sequences independent of their location and orientation. Local alignments are very powerful in detecting evolutionary rearrangements resulting in DNA reshuffling and segmental duplications (paralogs) as well as species-specific tandem gene expansions. In addition, local alignment tools are also useful when highly divergent genomes are compared, since gene structure and order is not well preserved over large evolutionary distances.

The first available alignment tools were designed to recognize and align highly homologous protein sequences. The basic local alignment search tool (BLAST) was created to rapidly match a relatively short stretch of DNA with homologous regions from a large collection of sequences stored in the National Center for Biotechnology Information (NCBI) database. Most recently, BLAST has evolved into a family of alignment tools able to detect matches for various types of sequences and evolutionary distances including blastn (nucleotide), blastp (protein), blastx [nucleotide query—protein database (db)], tblastn (protein query—translated db), tblastx (nucleotide query—translated db), and megablast (highly conserved matches). The “Blast2sequences” (bl2seq) tool was created to apply all the BLASTs to both nucleotide and protein pairwise sequence comparisons and is extremely powerful in annotating genomic sequences by comparing large contigs with mRNA sequences (Altschul *et al.*, 1990, 1997). Despite the great versatility of the BLAST, its application becomes limited when trying to align large genomic loci, megabases in length. Processing large alignments require graphical interfaces that allow the compact visualization of genes and repetitive elements along with the evolutionarily conservation profile of the aligned sequences.

A new generation of alignment tools can efficiently process two or more input sequences that can be up to genome size in length, are publicly accessible, and have user-friendly web interfaces. Some examples are listed in Table 10.2. PipMaker (Schwartz *et al.*, 2000), zPicture (Ovcharenko *et al.*, 2004a), and Mulan (Ovcharenko *et al.*, 2005) are based on the BLASTZ (Schwartz *et al.*,

**Table 10.2.** Comparative Genomic Tools

| Alignment tools | Web address   | Generated alignments                                   |
|-----------------|---|--|
| Alfresco        | <a href="http://www.sanger.ac.uk/Software/Alfresco/">http://www.sanger.ac.uk/Software/Alfresco/</a>                 | M  |
| AVID            | <a href="http://baboon.math.berkeley.edu/mavid/">http://baboon.math.berkeley.edu/mavid/</a>                         | G, M   |
| BALSA           | <a href="http://bayesweb.wadsworth.org/balsa/balsa.html">http://bayesweb.wadsworth.org/balsa/balsa.html</a>         | L  |
| BLAST           | <a href="http://www.ncbi.nlm.nih.gov/BLAST/">http://www.ncbi.nlm.nih.gov/BLAST/</a>                                 | L  |
| Blast2Sequences | <a href="http://www.ncbi.nlm.nih.gov/blast/bl2seq/bl2.html">http://www.ncbi.nlm.nih.gov/blast/bl2seq/bl2.html</a>   | L  |
| BLAT            | <a href="http://genome.ucsc.edu/cgi-bin/hgBlat">http://genome.ucsc.edu/cgi-bin/hgBlat</a>                           | L  |
| ClustalW        | <a href="http://www.ebi.ac.uk/clustalw/">http://www.ebi.ac.uk/clustalw/</a>   | M  |
| EMBOSS          | <a href="http://www.ebi.ac.uk/emboss/align/">http://www.ebi.ac.uk/emboss/align/</a>                                 | L, G   |
| GLASS           | <a href="http://glass.lcs.mit.edu/">http://glass.lcs.mit.edu/</a>   |  |
| LAGAN           | <a href="http://lagan.stanford.edu/">http://lagan.stanford.edu/</a>   | L, G, M  |
| LALIGN          | <a href="http://www.ch.embnet.org/software/LALIGN_form.html">http://www.ch.embnet.org/software/LALIGN_form.html</a> | L, G   |
| Mulan           | <a href="http://mulan.dcode.org">http://mulan.dcode.org</a>   | L, M   |
| MUMmer          | <a href="http://www.tigr.org/software/mummer/">http://www.tigr.org/software/mummer/</a>                             |  |
| Pfam            | <a href="http://www.sanger.ac.uk/Software/Pfam/">http://www.sanger.ac.uk/Software/Pfam/</a>                         | M  |
| PipMaker        | <a href="http://bio.cse.psu.edu/pipmaker/">http://bio.cse.psu.edu/pipmaker/</a>                                     | L, M   |
| Poa             | <a href="http://www.bioinformatics.ucla.edu/poa/">http://www.bioinformatics.ucla.edu/poa/</a>                       | M  |
| SAM             | <a href="http://www.cse.ucsc.edu/research/compbio/sam.html">http://www.cse.ucsc.edu/research/compbio/sam.html</a>   | HMM, M   |
| SSAHA           | <a href="http://www.sanger.ac.uk/Software/analysis/SSAHA/">http://www.sanger.ac.uk/Software/analysis/SSAHA/</a>     |  |
| SSEARCH         | <a href="http://www.biology.wustl.edu/gcg/ssearch.html">http://www.biology.wustl.edu/gcg/ssearch.html</a>           |  |
| SynPlot         | <a href="http://www.sanger.ac.uk/Users/jgrg/SynPlot">http://www.sanger.ac.uk/Users/jgrg/SynPlot</a>                 |  |
| Tcoffee         | <a href="http://www.ch.embnet.org/software/TCoffee.html">http://www.ch.embnet.org/software/TCoffee.html</a>         |  |
| VISTA           | <a href="http://www-gsd.lbl.gov/VISTA/index.html">http://www-gsd.lbl.gov/VISTA/index.html</a>                       | G, M   |
| zPicture        | <a href="http://zpicture.dcode.org">http://zpicture.dcode.org</a>   | L  |
| WUBlast2        | <a href="http://www.ebi.ac.uk/blast2/">http://www.ebi.ac.uk/blast2/</a>   | L  |
| Genome Browsers |   |  |
| ECR Browser     | <a href="http://ecrbrowser.dcode.org/">http://ecrbrowser.dcode.org/</a>   | (Ovcharenko <i>et al.</i> , 2004b)                     |
| Ensembl         | <a href="http://www.ensembl.org/">http://www.ensembl.org/</a>   | (Stalker <i>et al.</i> , 2004)                         |
| UCSC            | <a href="http://www.genome.ucsc.edu/">http://www.genome.ucsc.edu/</a>   | (Kent <i>et al.</i> , 2002; Kuhn <i>et al.</i> , 2007) |
| VISTA           | <a href="http://pipeline.lbl.gov/">http://pipeline.lbl.gov/</a>   | (Brudno <i>et al.</i> , 2004)                          |

G, global; M, multiple; L, local alignments; HMM, hidden Markov model. This is not meant to be a comprehensive list but a sampling of available resources.

2003) local alignment program. They combine suffix tree algorithms with dynamic programming techniques and have the capacity to align very long genomic intervals in a very short period of time. VISTA (Frazer *et al.*, 2004; Mayor *et al.*, 2000) is a visualization tool for global alignments generated by AVID (Bray *et al.*, 2003) or LAGAN (Brudno *et al.*, 2003) programs. All these alignment engines provide the user with informative, high-resolution graphical displays of the resulting alignments depicting both the genomic location of the conserved regions in the reference sequence and the degree of similarity for each aligned DNA segment. Individual features of the DNA sequence, such as coding exons, untranslated regions (UTR), and repetitive elements, can be distinguished in the graphical output through the use of different color schemes, allowing the identification of evolutionarily conserved sequences present in noncoding regions.

Comparative sequence alignment data can also be retrieved from genome browsers. Genome browsers are web-based database interfaces designed to allow the navigation across an entire genome by scrolling and zooming through any region of DNA and visualizing all available annotation data. In general, annotations include mRNAs, expressed sequence tags (EST), gene predictions, single nucleotide polymorphisms (SNPs), as well as many other features. Users can enter a region of the genome by searching for a landmark such as the name or acronym of a known gene, the accession number of a DNA sequence, the numeric position within a chromosome, or even through a homology search by providing a piece of sequence. The two main genome browsers are Ensembl (Stalker *et al.*, 2004) and the UCSC Genome Browser database (Kent *et al.*, 2002; Kuhn *et al.*, 2007), both of which were originally designed to support the assembly and annotation needs of the Human Genome Project by creating an efficient, user-friendly data storage and retrieval system with a compact visual presentation. The UCSC Browser has rapidly expanded to provide access to other available genome assemblies and their accompanying annotations, which now include 14 vertebrate species. Recently, UCSC Browser has also incorporated comparative genomic tracks to visualize regions of DNA conservation between two fully sequenced genomes, as well as a regulatory potential track based on normalized log-odds scores calculated using an HMM model that distinguishes known regulatory regions from ancestral repeats. Similar pairwise whole-genome alignments have been generated using a combination of local and global alignment strategies and can be visualized in the ECR Browser (Ovcharenko *et al.*, 2004b) and in the VISTA Genome Browser (Brudno *et al.*, 2004). In addition to ready-made pairwise alignments, the ECR Browser also aligns user-provided sequences to available genomes and incorporates any available annotation into the visual display of the generated alignment. For a comprehensive review of genome browsers and databases, see Ureta-Vidal *et al.* (2003).

## B. Predicting TFBSs

In eukaryotes, modulation of gene expression is achieved through the complex interaction of regulatory proteins (*trans*-factors) with specific DNA regions (*cis*-acting regulatory sequences). Intensive efforts over the last decades have identified numerous regulatory proteins or TFs and the DNA sequences they recognize. TRANSFAC database (<http://www.biobase.de>) represent the most comprehensive collection of TF-binding specificities, summarized as position weight matrices (PWMs) (Heinemeyer *et al.*, 1998, 1999). Pattern-recognition programs such as MATCH or MatInspector (Quandt *et al.*, 1995) use these libraries of TF-PWMs to identify significant matches in DNA sequences. A major confounding factor in the use of PWMs to identify TFBSs is that TFs bind to short (6–12 bp), degenerate sequence motifs that occur very frequently in a genome, and only a small fraction of the predicted TFBSs are functionally significant.

It has been shown that by combining pattern recognition with comparative sequence analysis the number of false positives is dramatically reduced while the number of functional sites is preserved. These results suggest an alternative strategy for sequence-based discovery of biologically relevant regulatory elements. The rVISTA (Loots and Ovcharenko, 2004) and Consite (Sandelin *et al.*, 2004) web-based tools combine TFBS motif searches and cross-species sequence analysis. rVISTA analysis proceeds in four major steps: (1) identify TFBS matches in each individual sequence using PWM from TRANSFAC database, (2) detect and calculate the percent identity of each locally aligned TFBS, (3) select TFBSs present in regions of high DNA conservation, and (4) graphically display individual or clustered TFBSs (Loots and Ovcharenko, 2004).

*Phylogenetic footprinting* is a method for identifying highly conserved DNA motifs present in a multiple sequence alignment. It is usually performed by computing a global multiple alignment of three or more orthologous sequences, and by identifying regions of high conservation in the alignment. FootPrinter (Blanchette and Tompa, 2003; Fang and Blanchette, 2006), FOOTER (Corcoran *et al.*, 2005), TRES, and PhyloGibbs (Siddharthan *et al.*, 2005) are some of the algorithms available for generating motif predictions and reporting motif sequences with the lowest parsimony scores, calculated with respect to the phylogenetic tree relating the input sequences. A more successful recent approach to phylogenetic footprinting is to use motif discovery algorithms such as MEME (Bailey *et al.*, 2006). Programs like MEME neither take into account the phylogenetic relationship among the input sequences nor do they rely on precalculated PWM stored in a database; they treat input sequences individually and the patterns are learned through several rounds of ungapped local alignments. The sampled alignments are used to fit a set of weights and the best weights are used to define an alignment, similar to the Gibbs sampling method (Schug and Overton, 1997).

In eukaryotes, transcriptional gene regulation is directed by a cohort of several different TFs that cooperatively bind to clusters of TFBS known as gene CRMs. One of the main objectives of transcriptional genomics is to decode the structure of CRMs and distinguish between the footprints of functional TFBS from genomic intervals devoid of biological significance and determine which CRM structures confer which tissue specificity. Searches for clusters of multiple adjacent binding sites for regulatory proteins have been successful in analyzing regulatory regions involved in mammalian muscle (Wasserman and Fickett, 1998) and liver-specific gene expression (Krivan and Wasserman, 2001). MSCAN (Alkema *et al.*, 2004) and rVISTA (Loots and Ovcharenko, 2004) are two examples of web-based tools that allow users to search for clusters of *cis*-elements, either using precalculated matrices from the TRANSFAC database or using consensus sequences provided by the user. Using these tools one could search for regions of high density of repeated TFBS for a single or multi different TFs. Recently, a new tool has been made available, SYNOR, which allows users to search for any configurations of TFBS across the whole human genome to predict functional regions with a distinct TFBS profile (Ovcharenko and Nobrega, 2005).

A new generation of computational tools aimed at predicting tissue-specific regulatory elements are blending together three elements: (a) comparative sequence analysis, (b) TFBS analysis, and (c) microarray expression analysis. In this approach, coexpressed genes are mapped to a genome and their promoters and surrounding conserved noncoding regions are used to identify CRMs that are overrepresented in the data set when compared with the distribution of the same CRMs across the whole genome. Pilpel and colleagues proposed such a method for modeling transcription regulatory networks in complex eukaryotes by combining microarray expression data with insights from combinatorial structure of promoter regions (Pilpel *et al.*, 2001). They were able to show that it is possible to discover novel functional PWMs by identifying statistically significant synergistic motifs in promoters of coexpressed genes, using a process called transcription factor centric clustering (TFCC). TFCC strategies are designed to create an explicit link between CRMs and the TFs that bind to them (Zhu *et al.*, 2002). These methods permit the detection of enriched TFBSs that are used as a seed to bicluster genes and compare gene expression with TFs' distribution in a two-dimensional space. Sharan *et al.* (2003) built on Pilpel's strategy by analyzing humans–mouse conserved promoter elements of cell cycle and stress response-related genes. Their analyses revealed several clusters of TFBS specific to the coexpressed genes. The significance of such co-occurrences was statistically evaluated and showed direct correlation between the identified CRMs and the biologically validated target genes derived from the microarray expression data. They proceeded to incorporate this method into a publicly available software,

CREME, which can perform combinatorial cluster analysis, statistically evaluate the detected co-occurrences, and graphically display predicted CRMs in a browser (Sharan *et al.*, 2003, 2004).

---

## IV. *IN VIVO* VALIDATION AND CHARACTERIZATION OF TRANSCRIPTIONAL REGULATORY ELEMENTS

In general, computational predictions have strongly correlated with functionally characterized regulatory elements mostly because the training sets used for these analyses have been carefully chosen from biologically validated data sets. On the contrary, the majority of novel predictions have not yet been biologically confirmed, and the limitations of computational tools have not been carefully assessed. Functional characterization of noncoding sequences represents the largest bottleneck which prevents us from expanding the annotation of regulatory elements from small target regions to entire genomes. The field of *in silico* biology is still in its infancy, but is evolving at a fast pace, presenting researchers with new theoretical solutions for the analysis of noncoding sequences. The computationally derived regulatory predictions may not all be functionally significant at this point, but by centering biological focus on a handful of high-priority regions to be tested, computational tools have already surpassed expectations for identifying regulatory elements. In [Section IV.A](#) we will review several experimental approaches employed to validate and characterize predicted transcriptional regulatory elements.

### A. Enhancer validation using transient transgenesis

Almost two decades ago, Kothary *et al.* (1988) created a mouse heat shock 68 promoter (*hsp68*)  $\beta$ -galactosidase (*LacZ*) transgene (*hsp68-LacZ*) and generated several independent lines of transgenic mice aiming to study heat shock gene regulation *in vivo*. In six of these transgenic lines, the transgene was consistently silent until subjected to heat shock treatment; however, one line of transgenic mice expressed *LacZ* in a neural-specific pattern independent of heat shock. The transgene integrated into the gene responsible for *dystonia musculorum* (OMIM 113810) (Bressman, 2003), mutated the gene, and acquired its transcriptional profile. This study was able to show that the *hsp68* promoter which is normally silent at physiological temperature is able to activate transcription of a reporter gene in response to positive regulatory elements and therefore can be used to trap enhancers in mammalian genomes (Kothary *et al.*, 1988). It was not until the human and mouse genome projects were well underway that this transgene was fully exploited and transformed into an essential tool for validating tissue-specific enhancer elements in transient transgenic mice.

Unlike stable transgenic lines that are screened for germline transmission of exogenous DNA, transient transgenic mice are transgenic animals that are analyzed in F0, without having to pass the transgene to future generations. In this method, embryos are injected with the reporter construct, transferred to the recipient mom, allowed to develop to a desired embryonic stage (usually between E10.5 and E14.5) when the moms are sacrificed and the embryos are harvested and examined for reporter transgene expression. This method dramatically shortens the experimental time for collecting expression data, since founder mice do not need to be established before carrying out the expression analysis, making this procedure highly efficient for validating a set of putative enhancers at a desired developmental time point. It is also a more cost-efficient approach, since it eliminates the need for breeding mice and establishing founder lines. The use of this method as a validation and characterization tool has dramatically grown over the past years, but in general has remained gene centric, where individual investigators have focused on testing conserved elements in the context of a well-characterized locus to identify tissue-specific enhancers that follow the expression pattern of one gene of interest (Bejerano *et al.*, 2006; Forghani *et al.*, 2001; Loots *et al.*, 2005; Nobrega *et al.*, 2003; Rojas *et al.*, 2005; Wang *et al.*, 2001; Zhu *et al.*, 2004). In an attempt to apply this method to more comprehensive, whole-genome analysis, this approach has proven very useful in validating highly conserved human elements grouped into two major categories: deeply conserved (evolutionarily conserved from human to fish) (Nobrega *et al.*, 2003; Pennacchio *et al.*, 2006) and ultraconserved (elements that have close to 100% identity from human to mouse) (Bejerano *et al.*, 2006; Pennacchio *et al.*, 2006; Poulin *et al.*, 2005).

While this method is currently considered *high throughput* in mice, there are several obstacles that preclude it from effectively being applied on a genome-wide scale. First, since mouse embryos develop *in utero*, collecting transgenic embryos is a terminal procedure that permits a litter of mice to be analyzed at only one given time point. In the absence of gene expression information for the genes putative enhancers are expected to regulate, screening for enhancer function could become a fishing expedition with low probability of success. For example, if an enhancer drives expression only at E17.5 in the medulla oblongata, with no detectable expression at any other time or tissue during development, the investigator would have to assay this particular time point to be able to detect its function. By assaying any other time point, the consistent lack of expression would drive the investigator to erroneously assume that the element has no function. A second drawback is posed by the transgene visualization method for *LacZ*. *LacZ* is a bacterial gene whose gene product, galactosidase, catalyzes the hydrolysis of galactosides or X-gal, and produces a blue color that can be visualized. This procedure requires fixation, and hence is terminal. Other caveats to this experimental approach include position effect, promoter specificity, and restriction to enhancer detection (one cannot detect a repressor or silencer element).

To overcome some of these problems, some alternatives have been proposed that include (1) the use of a transgenic model system that develops *ex utero* [fish (zebra fish) or frogs (*Xenopus*)]; (2) the use of reporter genes that do not require terminal fixation for visualization, such as green fluorescent protein (GFP); and (3) the use of larger transgene and “knocked” in reporters that track the protein expression from the endogenous promoter. Such transgenic systems would allow investigators to more efficiently monitor the transgene expression during development and determine both the temporal and spatial window of enhancer activity an element may pose. Using zebra fish transgenesis, Woolfe and his colleagues tested 25 ECRs from a set of 1400 elements identified by comparisons between the human and the puffer fish *Fugu rubripes* genomes. This study is of great significance because it makes two very important points. First, the authors were able to show that distant comparisons between human and *Fugu* enrich for a special category of regulatory elements which cluster around genes that have vital roles during embryonic development (e.g., TFs). Second, they were able to confirm that 23 of the total 25 tested ncECRs exhibit tissue-specific enhancer activity, suggesting that most of deeply conserved elements do indeed function as transcriptional regulatory elements (Woolfe *et al.*, 2005).

Recently, several reports have emerged that describe transposon-based gene delivery methods in both zebra fish and *Xenopus* embryos. These technologies can potentially evolve into a rapid system for transgenesis and expedite enhancer validation. In this approach, an ECR is cloned upstream of a minimal promoter driving a fluorescent reporter, where the entire transgenic cassette is flanked by transposable elements, such as *Tol2* (Allende *et al.*, 2006; Balciunas *et al.*, 2006; Hamlet *et al.*, 2006), *Sleeping Beauty* (Sinzelle *et al.*, 2006), *piggyBAC* (Wu *et al.*, 2006), or *Frog Prince* (Miskey *et al.*, 2003), and the reporter constructs are assayed for *cis*-regulatory activity. Using the *Tol2* transposable system in both zebra fish and *Xenopus* transgenic experiments, Allende and his colleagues have recently shown that most of the 50 ECRs they tested behave as positive modulators of gene expression and contribute to the specific temporal and spatial expression patterns of the endogenous genes they regulate (Allende *et al.*, 2006). The continuation of studies such as the ones described above will further our understanding of tissue-specific transcriptional regulation and will aid the discovery of all functional regulatory elements in the human genome.

## B. Identifying distant enhancers using large genomic constructs

An alternative approach to the use of heterologous minimal promoters to assay for enhancer activity in transient transgenics is to tag a transcript with a reporter gene (*LacZ* or *GFP*) within the context of a larger genomic region by modifying a yeast artificial chromosome (YAC) or BAC. This method has several advantages. First, the reporter gene is driven by the endogenous promoter and responds

to all regulatory elements included in the transgenic construct; therefore, by comparing the expression pattern of the reporter transgene to the endogenous expression pattern of the mouse gene one can determine what components of the complete tissue-specific expression profile is controlled by elements residing within the BAC region (Bouchard *et al.*, 2005; Gebhard *et al.*, 2007; Mortlock *et al.*, 2003; Tallini *et al.*, 2006). Second, transgenic animals generated using larger DNA constructs are less likely to be affected by position effects; therefore, most of the time the transgenic expression faithfully resembles the endogenous gene expression (Gebhard *et al.*, 2007). Third, by mutating individual ECRs within a BAC construct, one can identify not only regulatory elements that positively modulate transcription but also *cis*-elements that act as negative regulators, such as repressors, silencers, or boundary elements.

### C. Mutating candidate regulatory elements in engineered mice

The methods described in the previous sections are considered “gain-of-function” approaches to determining whether a conserved noncoding sequence possesses biological activity. However, these experiments do not provide any information whether these elements are functionally critical, and whether mutating them can lead to serious congenital abnormalities or whether they contribute to susceptibility to disease. The ultimate functional test that confirms essential physiological activity of a ncECR is to mutate these elements by changing individual base pairs or by removing them from an animal’s genome. Loss of function alleles can be generated by two main methods: random mutagenesis or targeted knock-out (KO). In a random mutagenesis experiment, one would subject an animal to a mutagen that either causes large chromosomal abnormalities, such as deletions and translocations, or has smaller effects by mutating single base pairs or removing a few nucleotides. Mutagenesis experiments are feasible for most experimental organisms, but require rigorous screenings to detect individuals that carry a desired mutation. A targeted mutation can only be engineered in animals for which KO technologies have been established, but unfortunately rodents are the only mammals for which targeted deletions can be carried out efficiently, primarily because embryonic stem cell lines have not been derived for other mammals. In this section, we will discuss functionally characterizing putative regulatory elements through loss-of-function experiments *in vivo*, in genetically modified mice by either removing an ECR from a large transgene or removing it from the mouse genome.

Homologous recombination techniques in yeast and bacteria have facilitated the genetic modification of artificial chromosomes [YACs, PI-based artificial chromosomes (PACs), and BACs] (Imam *et al.*, 2000; Lee *et al.*, 2001; Loots, 2006; Nistala and Sigmund, 2002; Warming *et al.*, 2005) to study gene expression and transcriptional regulation across large genomic loci.

The attraction of using such DNA constructs is primarily because they carry large fragments of genomic DNA (>100 kb) and therefore are likely to contain most of the *cis*-regulatory elements required for the expression of a gene; therefore, even when inserted randomly into the mouse genome, these transgenes are likely to behave similarly to their native environment recapitulating the gene expression pattern of endogenous loci. An extra advantage is our ability to modify these transgenes by inserting sequences prone to recombination such as *loxP* or *FRT* sites in the presence of *recombinases*. By flanking a ncECR with *loxP* sites, one can determine *in vivo* the transcriptional effects an element will have on a transcript when the ncECR is present or absent from the locus (Loots *et al.*, 2000), independent of position effects. In most situations, the integrated *loxP* sites do not affect gene expression and the floxed allele behaves equivalent to the wild-type allele. Upon administration of recombinase protein, the *loxP* sites excise the ncECR element, leaving behind a new deleted allele. Finally the investigator compares expression of the transgene with and without the ncECR to determine if the ncECR has any impact on transcription. This method was used to show that a highly conserved noncoding DNA sequence controls the expression of three cytokine genes, *IL4*, *IL13*, and *IL5*, in the context of a human YAC (Loots *et al.*, 2000). Similarly, one can use *in vitro* recombination to create several variants of a transgene, by either removing a putative regulatory element (Xu *et al.*, 2006) or removing a large noncoding region (Loots *et al.*, 2005). Since each transgene randomly integrates into the mouse genome, to ensure that an observed difference in expression is due to the mutation and not to position effect, several independent transgenic lines have to be analyzed for each allele. Finally, the most informative and reliable method of testing whether a ncECR impacts gene expression and causes a deleterious phenotype is to remove it from the mouse genome through targeted KO strategies. Although this approach remains the ultimate proof of biological significance because it is technically challenging, laborious, expensive, and time-consuming to generate KO animals, to date very few ncECR have been mutated in mice (Mohrs *et al.*, 2001; Sagai *et al.*, 2005).

---

## V. CONCLUSIONS

Comparative genomic approaches are having a remarkable impact on the study of transcriptional regulation in eukaryotes. Many eukaryotic genome sequences are being explored by new computational methods and high-throughput experimental tools. These tools are enabling efficient searches for common regulatory motifs which will eventually lead to elucidating the genome's second code: understanding the building blocks of tissue-specific gene regulation encoded in noncoding DNA. Experimental validation and characterization, however, continues to be a major bottleneck, and hence extending the limits of current

techniques will greatly enhance the discovery of transcriptional regulatory elements in mammals, moving us closer to a systematic deciphering of transcriptional regulatory elements and providing the first global insights into gene regulatory networks. In addition to the methods described here, other recent advances in transcriptional regulation approaches that include probing DNA–protein interactions by ChIP-chip or comparing patterns of gene expression are moving the field of transcriptional genomics forward.

## References

- Ahituv, N., Rubin, E. M., and Nobrega, M. A. (2004). Exploiting human–fish genome comparisons for deciphering gene regulation. *Hum. Mol. Genet.* **13**(Spec. No. 2), R261–R266.
- Alkema, W. B., Johansson, O., Lagergren, J., and Wasserman, W. W. (2004). MSCAN: Identification of functional clusters of transcription factor binding sites. *Nucleic Acids Res.* **32**, W195–W198.
- Allende, M. L., Manzanares, M., Tena, J. J., Feijoo, C. G., and Gomez-Skarmeta, J. L. (2006). Cracking the genome's second code: Enhancer detection by combined phylogenetic footprinting and transgenic fish and frog embryos. *Methods* **39**, 212–219.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
- Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402.
- Bailey, T. L., Williams, N., Misleh, C., and Li, W. W. (2006). MEME: Discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.* **34**, W369–W373.
- Balciunas, D., Wangenstein, K. J., Wilber, A., Bell, J., Geurts, A., Sivasubbu, S., Wang, X., Hackett, P. B., Largaespada, D. A., McIvor, R. S., and Ekker, S. C. (2006). Harnessing a high cargo-capacity transposon for genetic applications in vertebrates. *PLoS Genet.* **2**, 1715–1724.
- Balemans, W., Ebeling, M., Patel, N., Van Hul, E., Olson, P., Dioszegi, M., Lacza, C., Wuys, W., Van Den Ende, J., Willems, P., Paes-Alves, A. F., Hill, S., et al. (2001). Increased bone density in sclerosteosis is due to the deficiency of a novel secreted protein (SOST). *Hum. Mol. Genet.* **10**, 537–543.
- Balemans, W., Patel, N., Ebeling, M., Van Hul, E., Wuys, W., Lacza, C., Dioszegi, M., Dikkers, F. G., Hilderling, P., Willems, P. J., Verheij, J. B., Lindpaintner, K., et al. (2002). Identification of a 52 kb deletion downstream of the SOST gene in patients with van Buchem disease. *J. Med. Genet.* **39**, 91–97.
- Banet, G., Bibi, O., Matouk, I., Ayesh, S., Laster, M., Kimber, K. M., Tykocinski, M., de Groot, N., Hochberg, A., and Ohana, P. (2000). Characterization of human and mouse H19 regulatory sequences. *Mol. Biol. Rep.* **27**, 157–165.
- Bejerano, G., Lowe, C. B., Ahituv, N., King, B., Siepel, A., Salama, S. R., Rubin, E. M., Kent, W. J., and Haussler, D. (2006). A distal enhancer and an ultraconserved exon are derived from a novel retroposon. *Nature* **441**, 87–90.
- Berman, B. P., Nibu, Y., Pfeiffer, B. D., Tomancak, P., Celniker, S. E., Levine, M., Rubin, G. M., and Eisen, M. B. (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.
- Blanchette, M., and Tompa, M. (2003). FootPrinter: A program designed for phylogenetic footprinting. *Nucleic Acids Res.* **31**, 3840–3842.
- Blanchette, M., Schwikowski, B., and Tompa, M. (2002). Algorithms for phylogenetic footprinting. *J. Comput. Biol.* **9**, 211–223.

- Bouchard, M., Grote, D., Craven, S. E., Sun, Q., Steinlein, P., and Busslinger, M. (2005). Identification of Pax2-regulated genes by expression profiling of the mid-hindbrain organizer region. *Development* **132**, 2633–2643.
- Bray, N., Dubchak, I., and Pachter, L. (2003). AVID: A global alignment program. *Genome Res.* **13**, 97–102.
- Bressman, S. B. (2003). Dystonia: Phenotypes and genotypes. *Rev. Neurol. (Paris)* **159**, 849–856.
- Brudno, M., Do, C. B., Cooper, G. M., Kim, M. F., Davydov, E., Green, E. D., Sidow, A., and Batzoglou, S. (2003). LAGAN and Multi-LAGAN: Efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* **13**, 721–731.
- Brudno, M., Poliakov, A., Salamov, A., Cooper, G. M., Sidow, A., Rubin, E. M., Solovyev, V., Batzoglou, S., and Dubchak, I. (2004). Automated whole-genome multiple alignment of rat, mouse, and human. *Genome Res.* **14**, 685–692.
- Catena, R., Tiveron, C., Ronchi, A., Porta, S., Ferri, A., Tatangelo, L., Cavallaro, M., Favaro, R., Ottolenghi, S., Reinbold, R., Schöler, H., and Nicolis, S. K. (2004). Conserved POU binding DNA sites in the Sox2 upstream enhancer regulate gene expression in embryonic and neural stem cells. *J. Biol. Chem.* **279**, 41846–41857.
- Chi, X., Chatterjee, P. K., Wilson, W., 3rd, Zhang, S. X., Demayo, F. J., and 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.
- Chiang, D. Y., Moses, A. M., Kellis, M., Lander, E. S., and Eisen, M. B. (2003). Phylogenetically and spatially conserved word pairs associated with gene-expression changes in yeasts. *Genome Biol.* **4**, R43.
- Collins, F. (2004). Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945.
- Corcoran, D. L., Feingold, E., and Benos, P. V. (2005). FOOTER: A web tool for finding mammalian DNA regulatory regions using phylogenetic footprinting. *Nucleic Acids Res.* **33**, W442–W446.
- Crisponi, L., Uda, M., Deiana, M., Loi, A., Nagaraja, R., Chiappe, F., Schlessinger, D., Cao, A., and Pilia, G. (2004). FOXL2 inactivation by a translocation 171 kb away: Analysis of 500 kb of chromosome 3 for candidate long-range regulatory sequences. *Genomics* **83**, 757–764.
- Danielian, P. S., Echelard, Y., Vassileva, G., and McMahon, A. P. (1997). A 5.5-kb enhancer is both necessary and sufficient for regulation of Wnt-1 transcription in vivo. *Dev. Biol.* **192**, 300–309.
- de Kok, Y. J., Merks, G. F., van der Maarel, S. M., Huber, I., Malcolm, S., Ropers, H. H., and Cremers, F. P. (1995). A duplication/paracentric inversion associated with familial X-linked deafness (DFN3) suggests the presence of a regulatory element more than 400 kb upstream of the POU3F4 gene. *Hum. Mol. Genet.* **4**, 2145–2150.
- Fang, F., and Blanchette, M. (2006). FootPrinter3: Phylogenetic footprinting in partially alignable sequences. *Nucleic Acids Res.* **34**, W617–W620.
- Fang, J., Dagenais, S. L., Erickson, R. P., Arlt, M. F., Glynn, M. W., Gorski, J. L., Seaver, L. H., and Glover, T. W. (2000). Mutations in FOXC2 (MFH-1), a forkhead family transcription factor, are responsible for the hereditary lymphedema-distichiasis syndrome. *Am. J. Hum. Genet.* **67**, 1382–1388.
- Fantes, J., Redeker, B., Breen, M., Boyle, S., Brown, J., Fletcher, J., Jones, S., Bickmore, W., Fukushima, Y., Mannens, M., Danes, S., van Heyningen, V., et al. (1995). Aniridia-associated cytogenetic rearrangements suggest that a position effect may cause the mutant phenotype. *Hum. Mol. Genet.* **4**, 415–422.
- Fink, D. L., Chen, R. O., Noller, H. F., and Altman, R. B. (1996). Computational methods for defining the allowed conformational space of 16S rRNA based on chemical footprinting data. *RNA* **2**, 851–866.
- Flomen, R. H., Gorman, P. A., Vatcheva, R., Groet, J., Barisic, I., Ligutic, I., Sheer, D., and Nizetic, D. (1997). Rieger syndrome locus: A new reciprocal translocation t(4;12)(q25;q15) and a deletion del(4)(q25q27) both break between markers D4S2945 and D4S193. *J. Med. Genet.* **34**, 191–195.

- Forghani, R., Garofalo, L., Foran, D. R., Farhadi, H. F., Lepage, P., Hudson, T. J., Tretjakoff, I., Valera, P., and Peterson, A. (2001). A distal upstream enhancer from the myelin basic protein gene regulates expression in myelin-forming schwann cells. *J. Neurosci.* **21**, 3780–3787.
- Frazer, K. A., Pachter, L., Poliakov, A., Rubin, E. M., and Dubchak, I. (2004). VISTA: Computational tools for comparative genomics. *Nucleic Acids Res.* **32**, W273–W279.
- Gebhard, S., Hattori, T., Bauer, E., Bosl, M. R., Schlund, B., Poschl, E., Adam, N., de Crombrughe, B., and von der Mark, K. (2007). BAC constructs in transgenic reporter mouse lines control efficient and specific LacZ expression in hypertrophic chondrocytes under the complete Col10a1 promoter. *Histochem. Cell Biol.* **127**, 183–194.
- Gilligan, P., Brenner, S., and Venkatesh, B. (2002). Fugu and human sequence comparison identifies novel human genes and conserved non-coding sequences. *Gene* **294**, 35–44.
- Gottgens, B., Barton, L. M., Gilbert, J. G., Bench, A. J., Sanchez, M. J., Bahn, S., Mistry, S., Grafham, D., McMurray, A., Vaudin, M., Amaya, E., Bentley, D. R., et al. (2000). Analysis of vertebrate SCL loci identifies conserved enhancers. *Nat. Biotechnol.* **18**, 181–186.
- Hamlet, M. R., Yergeau, D. A., Kuliye, E., Takeda, M., Taira, M., Kawakami, K., and Mead, P. E. (2006). Tol2 transposon-mediated transgenesis in *Xenopus tropicalis*. *Genesis* **44**, 438–445.
- Heinemeyer, T., Wingender, E., Reuter, I., Hermjakob, H., Kel, A. E., Kel, O. V., Ignatieva, E. V., Ananko, E. A., Podkolodnaya, O. A., Kolpakov, F. A., Podkolodny, N. L., and Kolchanov, N. A. (1998). Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. *Nucleic Acids Res.* **26**, 362–367.
- Heinemeyer, T., Chen, X., Karas, H., Kel, A. E., Kel, O. V., Liebich, I., Meinhardt, T., Reuter, I., Schacherer, F., and Wingender, E. (1999). Expanding the TRANSFAC database towards an expert system of regulatory molecular mechanisms. *Nucleic Acids Res.* **27**, 318–322.
- Hertz, G. Z., and Stormo, G. D. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15**, 563–577.
- Heutink, P., Zguricas, J., van Oosterhout, L., Breedveld, G. J., Testers, L., Sandkuijl, L. A., Snijders, P. J. L. M., Weissenbach, J., Lindhout, D., Hovius, S. E. R., and Oostra, B. A. (1994). The gene for triphalangeal thumb maps to the subtelomeric region of chromosome 7q. *Nat. Genet.* **6**, 287–292.
- Imam, A. M., Patrinos, G. P., de Krom, M., Bottardi, S., Janssens, R. J., Katsantoni, E., Wai, A. W., Sherratt, D. J., and Grosveld, F. G. (2000). Modification of human beta-globin locus PAC clones by homologous recombination in *Escherichia coli*. *Nucleic Acids Res.* **28**, E65.
- Inoue, K., Osaka, H., Thurston, V. C., Clarke, J. T., Yoneyama, A., Rosenbarker, L., Bird, T. D., Hodes, M. E., Shaffer, L. G., and Lupski, J. R. (2002). Genomic rearrangements resulting in PLP1 deletion occur by nonhomologous end joining and cause different dysmyelinating phenotypes in males and females. *Am. J. Hum. Genet.* **71**, 838–853.
- Jegga, A. G., Gupta, A., Gowrisankar, S., Deshmukh, M. A., Connolly, S., Finley, K., and Aronow, B. J. (2005). CisMols Analyzer: Identification of compositionally similar cis-element clusters in ortholog conserved regions of coordinately expressed genes. *Nucleic Acids Res.* **33**, W408–W411.
- Jegga, A. G., Chen, J., Gowrisankar, S., Deshmukh, M. A., Gudivada, R., Kong, S., Kaimal, V., and Aronow, B. J. (2007). GenomeTrafac: A whole genome resource for the detection of transcription factor binding site clusters associated with conventional and microRNA encoding genes conserved between mouse and human gene orthologs. *Nucleic Acids Res.* **35**, D116–D121.
- Kamat, A., Graves, K. H., Smith, M. E., Richardson, J. A., and Mendelson, C. R. (1999). A 500-bp region, approximately 40 kb upstream of the human CYP19 (aromatase) gene, mediates placenta-specific expression in transgenic mice. *Proc. Natl. Acad. Sci. USA* **96**, 4575–4580.
- Kent, W. J., and Zahler, A. M. (2000). Conservation, regulation, synteny, and introns in a large-scale *C. briggsae*-*C. elegans* genomic alignment. *Genome Res.* **10**, 1115–1125.
- Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002). The human genome browser at UCSC. *Genome Res.* **12**, 996–1006.

- Khokha, M. K., Hsu, D., Brunet, L. J., Dionne, M. S., and Harland, R. M. (2003). Gremlin is the BMP antagonist required for maintenance of Shh and Fgf signals during limb patterning. *Nat. Genet.* **34**, 303–307.
- Kim, J. D., Hinz, A. K., Bergmann, A., Huang, J. M., Ovcharenko, I., Stubbs, L., and Kim, J. (2006). Identification of clustered YY1 binding sites in imprinting control regions. *Genome Res.* **16**, 901–911.
- Kioussis, D., Vanin, E., deLange, T., Flavell, R. A., and Grosveld, F. G. (1983). Beta-globin gene inactivation by DNA translocation in gamma beta-thalassaemia. *Nature* **306**, 662–666.
- Kothary, R., Clapoff, S., Brown, A., Campbell, R., Peterson, A., and Rossant, J. (1988). A transgene containing lacZ inserted into the dystonia locus is expressed in neural tube. *Nature* **335**, 435–437.
- Krebsbach, P. H., Nakata, K., Bernier, S. M., Hatano, O., Miyashita, T., Rhodes, C. S., and Yamada, Y. (1996). Identification of a minimum enhancer sequence for the type II collagen gene reveals several core sequence motifs in common with the link protein gene. *J. Biol. Chem.* **271**, 4298–4303.
- Krivan, W., and Wasserman, W. W. (2001). A predictive model for regulatory sequences directing liver-specific transcription. *Genome Res.* **11**, 1559–1566.
- Kuhn, R. M., Karolchik, D., Zweig, A. S., Trumbower, H., Thomas, D. J., Thakkapallayil, A., Sugnet, C. W., Stanke, M., Smith, K. E., Siepel, A., Rosenbloom, K. R., Rhead, B., et al. (2007). The UCSC genome browser database: Update 2007. *Nucleic Acids Res.* **35**, D668–D673.
- Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.
- Lee, E. C., Yu, D., Martinez de Velasco, J., Tessarollo, L., Swing, D. A., Court, D. L., Jenkins, N. A., and Copeland, N. G. (2001). A highly efficient Escherichia coli-based chromosome engineering system adapted for recombinogenic targeting and subcloning of BAC DNA. *Genomics* **73**, 56–65.
- Lettice, L. A., Heaney, S. J., Purdie, L. A., Li, L., de Beer, P., Oostra, B. A., Goode, D., Elgar, G., Hill, R. E., and de Graaff, E. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.* **12**, 1725–1735.
- Levine, M., and Tjian, R. (2003). Transcription regulation and animal diversity. *Nature* **424**, 147–151.
- Loots, G. G. (2006). Modifying yeast artificial chromosomes to generate Cre/LoxP and FLP/FRT site-specific deletions and inversions. *Methods Mol. Biol.* **349**, 75–84.
- Loots, G. G., and Ovcharenko, I. (2004). rVISTA 2.0: Evolutionary analysis of transcription factor binding sites. *Nucleic Acids Res.* **32**, W217–W221.
- Loots, G. G., Locksley, R. M., Blankespoor, C. M., Wang, Z. E., Miller, W., Rubin, E. M., and Frazer, K. A. (2000). Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288**, 136–140.
- Loots, G. G., Ovcharenko, I., Pachter, L., Dubchak, I., and Rubin, E. M. (2002). rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res.* **12**, 832–839.
- Loots, G. G., Kneissel, M., Keller, H., Baptist, M., Chang, J., Collette, N. M., Ovcharenko, D., Plajzer-Frick, I., and Rubin, E. M. (2005). Genomic deletion of a long-range bone enhancer misregulates sclerostin in Van Buchem disease. *Genome Res.* **15**, 928–935.
- Majoros, W. H., Pertea, M., Delcher, A. L., and Salzberg, S. L. (2005). Efficient decoding algorithms for generalized hidden Markov model gene finders. *BMC Bioinformatics* **6**, 16.
- 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., et al. (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.

- Marshall, P., Chartrand, N., and Worton, R. G. (2001). The mouse dystrophin enhancer is regulated by MyoD, E-box-binding factors, and by the serum response factor. *J. Biol. Chem.* **276**, 20719–20726.
- Mass, R. L., Zeller, R., Woychik, R. P., Vogt, T. F., and Leder, P. (1990). Disruption of formin-encoding transcripts in two mutant limb deformity alleles. *Nature* **346**, 853–855.
- Masuya, H., Sagai, T., Wakana, S., Moriwaki, K., and Shiroishi, T. (1995). A duplicated zone of polarizing activity in polydactylous mouse mutants. *Genes Dev.* **9**, 1645–1653.
- Mayor, C., Brudno, M., Schwartz, J. R., Poliakov, A., Rubin, E. M., Frazer, K. A., Pachter, L. S., and Dubchak, I. (2000). VISTA: Visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16**, 1046–1047.
- McElreavy, K., Vilain, E., Abbas, N., Costa, J. M., Souleyreau, N., Kucheria, K., Boucekine, C., Thibaud, E., Brauner, R., Flamant, F., and Fellous, M. (1992). XY sex reversal associated with a deletion 5' to the SRY "HMG box" in the testis-determining region. *Proc. Natl. Acad. Sci. USA* **89**, 11016–11020.
- Miskey, C., Izsvak, Z., Plasterk, R. H., and Ivics, Z. (2003). The Frog Prince: A reconstructed transposon from *Rana pipiens* with high transpositional activity in vertebrate cells. *Nucleic Acids Res.* **31**, 6873–6881.
- Mohrs, M., Blankespoor, C. M., Wang, Z. E., Loots, G. G., Afzal, V., Hadeiba, H., Shinkai, K., Rubin, E. M., and Locksley, R. M. (2001). Deletion of a coordinate regulator of type 2 cytokine expression in mice. *Nat. Immunol.* **2**, 842–847.
- Mortlock, D. P., Guenther, C., and Kingsley, D. M. (2003). A general approach for identifying distant regulatory elements applied to the *Gdf6* gene. *Genome Res.* **13**, 2069–2081.
- Needleman, S. B., and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453.
- Nistala, R., and Sigmund, C. D. (2002). A reliable and efficient method for deleting operational sequences in PACs and BACs. *Nucleic Acids Res.* **30**, e41.
- Nobrega, M. A., Ovcharenko, I., Afzal, V., and Rubin, E. M. (2003). Scanning human gene deserts for long-range enhancers. *Science* **302**, 413.
- Ovcharenko, I., and Nobrega, M. A. (2005). Identifying synonymous regulatory elements in vertebrate genomes. *Nucleic Acids Res.* **33**, W403–W407.
- Ovcharenko, I., Loots, G. G., Hardison, R. C., Miller, W., and Stubbs, L. (2004a). zPicture: Dynamic alignment and visualization tool for analyzing conservation profiles. *Genome Res.* **14**, 472–477.
- Ovcharenko, I., Nobrega, M. A., Loots, G. G., and Stubbs, L. (2004b). ECR Browser: A tool for visualizing and accessing data from comparisons of multiple vertebrate genomes. *Nucleic Acids Res.* **32**, W280–W286.
- Ovcharenko, I., Loots, G. G., Giardine, B. M., Hou, M., Ma, J., Hardison, R. C., Stubbs, L., and Miller, W. (2005). Mulan: Multiple-sequence local alignment and visualization for studying function and evolution. *Genome Res.* **15**, 184–194.
- Pennacchio, L. A., Olivier, M., Hubacek, J. A., Cohen, J. C., Cox, D. R., Fruchart, J. C., Krauss, R. M., and Rubin, E. M. (2001). An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. *Science* **294**, 169–173.
- Pennacchio, L. A., Ahituv, N., Moses, A. M., Prabhakar, S., Nobrega, M. A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K. D., Plajzer-Frick, I., Akiyama, J., et al. (2006). In vivo enhancer analysis of human conserved non-coding sequences. *Nature* **444**, 499–502.
- Pennacchio, L. A., Loots, G. G., Nobrega, M. A., and Ovcharenko, I. (2007). Predicting tissue-specific enhancers in the human genome. *Genome Res.* **17**, 201–211.
- Pilpel, Y., Sudarsanam, P., and Church, G. M. (2001). Identifying regulatory networks by combinatorial analysis of promoter elements. *Nat. Genet.* **29**, 153–159.
- Poulin, F., Nobrega, M. A., Plajzer-Frick, I., Holt, A., Afzal, V., Rubin, E. M., and Pennacchio, L. A. (2005). In vivo characterization of a vertebrate ultraconserved enhancer. *Genomics* **85**, 774–781.

- Quandt, K., Frech, K., Karas, H., Wingender, E., and Werner, T. (1995). MatInd and MatInspector: New fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.* **23**, 4878–4884.
- Roessler, E., Ward, D. E., Gaudenz, K., Belloni, E., Scherer, S. W., Donnai, D., Siegel-Bartelt, J., Tsui, L. C., and Muenke, M. (1997). Cytogenetic rearrangements involving the loss of the Sonic Hedgehog gene at 7q36 cause holoprosencephaly. *Hum. Genet.* **100**, 172–181.
- Rojas, A., De Val, S., Heidt, A. B., Xu, S. M., Bristow, J., and Black, B. L. (2005). Gata4 expression in lateral mesoderm is downstream of BMP4 and is activated directly by Forkhead and GATA transcription factors through a distal enhancer element. *Development* **132**, 3405–3417.
- Sagai, T., Hosoya, M., Mizushina, Y., Tamura, M., and Shiroishi, T. (2005). Elimination of a long-range cis-regulatory module causes complete loss of limb-specific Shh expression and truncation of the mouse limb. *Development* **132**, 797–803.
- Sandelin, A., Wasserman, W. W., and Lenhard, B. (2004). ConSite: Web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Res.* **32**, W249–W252.
- Schug, J., and Overton, G. C. (1997). Modeling transcription factor binding sites with Gibbs Sampling and Minimum Description Length encoding. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **5**, 268–271.
- Schwartz, S., Zhang, Z., Frazer, K. A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R., and Miller, W. (2000). PipMaker—a web server for aligning two genomic DNA sequences. *Genome Res.* **10**, 577–586.
- Schwartz, S., Kent, W. J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R. C., Haussler, D., and Miller, W. (2003). Human-mouse alignments with BLASTZ. *Genome Res.* **13**, 103–107.
- Sharan, R., Ovcharenko, I., Ben-Hur, A., and Karp, R. M. (2003). CREME: A framework for identifying cis-regulatory modules in human-mouse conserved segments. *Bioinformatics* **19** (Suppl. 1), i283–i291.
- Sharan, R., Ben-Hur, A., Loots, G. G., and Ovcharenko, I. (2004). CREME: Cis-Regulatory Module Explorer for the human genome. *Nucleic Acids Res.* **32**, W253–W256.
- Siddharthan, R., Siggia, E. D., and van Nimwegen, E. (2005). PhyloGibbs: A Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput. Biol.* **1**, e67.
- Sinzelle, L., Vallin, J., Coen, L., Chesneau, A., Pasquier, D. D., Pollet, N., Demeneix, B., and Mazabraud, A. (2006). Generation of transgenic *Xenopus laevis* using the Sleeping Beauty transposon system. *Transgenic Res.* **15**, 751–760.
- Stalker, J., Gibbins, B., Meidl, P., Smith, J., Spooner, W., Hotz, H. R., and Cox, A. V. (2004). The Ensembl Web site: Mechanics of a genome browser. *Genome Res.* **14**, 951–955.
- Tallini, Y. N., Shui, B., Greene, K. S., Deng, K. Y., Doran, R., Fisher, P. J., Zipfel, W., and Kotlikoff, M. I. (2006). BAC transgenic mice express enhanced green fluorescent protein in central and peripheral cholinergic neurons. *Physiol. Genomics* **27**, 391–397.
- Thomas, M. C., and Chiang, C. M. (2006). The general transcription machinery and general cofactors. *Crit. Rev. Biochem. Mol. Biol.* **41**, 105–178.
- Touchman, J. W., Anikster, Y., Dietrich, N. L., Maduro, V. V., McDowell, G., Shotelersuk, V., Bouffard, G. G., Beckstrom-Sternberg, S. M., Gahl, W. A., and Green, E. D. (2000). The genomic region encompassing the nephropathic cystinosis gene (CTNS): Complete sequencing of a 200-kb segment and discovery of a novel gene within the common cystinosis-causing deletion. *Genome Res.* **10**, 165–173.
- Ureta-Vidal, A., Ettwiller, L., and Birney, E. (2003). Comparative genomics: Genome-wide analysis in metazoan eukaryotes. *Nat. Rev. Genet.* **4**, 251–262.
- van Deutekom, J. C., Lemmers, R. J., Grewal, P. K., van Geel, M., Romberg, S., Dauwerse, H. G., Wright, T. J., Padberg, G. W., Hofker, M. H., Hewitt, J. E., and Frants, R. R. (1996). Identification of the first gene (FRG1) from the FSHD region on human chromosome 4q35. *Hum. Mol. Genet.* **5**, 581–590.

- Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., *et al.* (2001). The sequence of the human genome. *Science* **291**, 1304–1351.
- Vortkamp, A., Gessler, M., and Grzeschik, K. H. (1991). GLI3 zinc-finger gene interrupted by translocations in Greig syndrome families. *Nature* **352**, 539–540.
- Wallis, D. E., Roessler, E., Hehr, U., Nanni, L., Wiltshire, T., Richieri-Costa, A., Gillissen-Kaesbach, G., Zackai, E. H., Rommens, J., and Muenke, M. (1999). Mutations in the homeodomain of the human SIX3 gene cause holoprosencephaly. *Nat. Genet.* **22**, 196–198.
- Wang, D. Z., Valdez, M. R., McAnally, J., Richardson, J., and Olson, E. N. (2001). The Mef2c gene is a direct transcriptional target of myogenic bHLH and MEF2 proteins during skeletal muscle development. *Development* **128**, 4623–4633.
- Warming, S., Costantino, N., Court, D. L., Jenkins, N. A., and Copeland, N. G. (2005). Simple and highly efficient BAC recombineering using galK selection. *Nucleic Acids Res.* **33**, e36.
- Wasserman, W. W., and Fickett, J. W. (1998). Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.* **278**, 167–181.
- Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S. E., Attwood, J., *et al.* (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562.
- Wirth, J., Wagner, T., Meyer, J., Pfeiffer, R. A., Tietze, H. U., Schempp, W., and Scherer, G. (1996). Translocation breakpoints in three patients with campomelic dysplasia and autosomal sex reversal map more than 130 kb from SOX9. *Hum. Genet.* **97**, 186–193.
- Woolfe, A., Goodson, M., Goode, D. K., Snell, P., McEwen, G. K., Vavouri, T., Smith, S. F., North, P., Callaway, H., Kelly, K., Walter, K., Abnizova, I., *et al.* (2005). Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.* **3**, e7.
- Woychik, R. P., and Alagramam, K. (1998). Insertional mutagenesis in transgenic mice generated by the pronuclear microinjection procedure. *Int. J. Dev. Biol.* **42**, 1009–1017.
- Woychik, R. P., Stewart, T. A., Davis, L. G., D'Eustachio, P., and Leder, P. (1985). An inherited limb deformity created by insertional mutagenesis in a transgenic mouse. *Nature* **318**, 36–40.
- Woychik, R. P., Generoso, W. M., Russell, L. B., Cain, K. T., Cacheiro, N. L., Bultman, S. J., Selby, P. B., Dickinson, M. E., Hogan, B. L., and Rutledge, J. C. (1990). Molecular and genetic characterization of a radiation-induced structural rearrangement in mouse chromosome 2 causing mutations at the limb deformity and agouti loci. *Proc. Natl. Acad. Sci. USA* **87**, 2588–2592.
- Wu, S. C., Meir, Y. J., Coates, C. J., Handler, A. M., Pelczar, P., Moisyadi, S., and Kaminski, J. M. (2006). piggyBac is a flexible and highly active transposon as compared to sleeping beauty, Tol2, and Mos1 in mammalian cells. *Proc. Natl. Acad. Sci. USA* **103**, 15008–15013.
- Xu, X., Scott, M. M., and 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.
- Zhu, Z., Pilpel, Y., and Church, G. M. (2002). Computational identification of transcription factor binding sites via a transcription-factor-centric clustering (TFCC) algorithm. *J. Mol. Biol.* **318**, 71–81.
- Zhu, L., Lee, H. O., Jordan, C. S., Cantrell, V. A., Southard-Smith, E. M., and Shin, M. K. (2004). Spatiotemporal regulation of endothelin receptor-B by SOX10 in neural crest-derived enteric neuron precursors. *Nat. Genet.* **36**, 732–737.
- Zuniga, A., Michos, O., Spitz, F., Haramis, A. P., Panman, L., Galli, A., Vintersten, K., Klasen, C., Mansfield, W., Kuc, S., Duboule, D., Dono, R., *et al.* (2004). Mouse limb deformity mutations disrupt a global control region within the large regulatory landscape required for Gremlin expression. *Genes Dev.* **18**, 1553–1564.