GENOMIC STRATEGIES TO IDENTIFY MAMMALIAN REGULATORY SEQUENCES

Len A. Pennacchio and Edward M. Rubin

With the continuing accomplishments of the human genome project, high-throughput strategies to identify DNA sequences that are important in mammalian gene regulation are becoming increasingly feasible. In contrast to the historic, labour-intensive, wet-laboratory methods for identifying regulatory sequences, many modern approaches are heavily focused on the computational analysis of large genomic data sets. Data from inter-species genomic sequence comparisons and genome-wide expression profiling, integrated with various computational tools, are poised to contribute to the decoding of genomic sequence and to the identification of those sequences that orchestrate gene regulation. In this review, we highlight several genomic approaches that are being used to identify regulatory sequences in mammalian genomes.

DNASEI HYPERSENSITIVITY ASSAY
Identifies regions of the genome that lack nucleosome structure and are therefore readily degraded by the enzyme DNaseI. Such regions tend to be associated with transcriptional activity.

DNA FOOTPRINTING ASSAY
An assay that identifies a region of DNA that is protected from digestion by DNaseI (usually due to the binding of a protein, such as a transcription factor).

Genome Sciences Department, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, California 94720, USA. Correspondence to E.R. e-mail: EMRubin@lbl.gov

Regulatory sequences constitute a small fraction of the roughly 95% of the mammalian genome that does not encode proteins, but they determine the level, location and chronology of gene expression. Despite the importance of these non-coding sequences in gene regulation, our ability to identify and predict functions for this category of DNA is extremely limited. This contrasts with coding sequences, where the long-term availability and study of cDNAs and proteins has contributed to the development of a defined vocabulary that allows their identity and their function to be predicted, in many cases from the analysis of sequence alone. Because of the paucity of information about non-coding sequences involved in gene regulation, few alterations of these sequences have been linked to disease susceptibility, despite their predicted role in many common human disorders.

Classical searches for *cis*-regulatory sequences (BOX 1) have typically involved various trial-and-error strategies. The focus on the identification of regulatory elements for individual genes has included several experimental approaches: the generation of deletion constructs to determine the minimal sequences necessary for transcription in cell-culture-based systems; DNASE I HYPERSENSITIVITY STUDIES to identify sequences potentially available for transcription factor binding; and *in vitro* approaches, such as DNA FOOTPRINTING and GEL SHIFTS, to determine sequences that bind various regulatory proteins. Screens to identify *cis*-regulatory elements have also been carried out in transgenic mice, albeit in an extremely laborious and low-throughput manner. In addition, a limited number of large-scale promoter and enhancer trapping studies have been done[1–3]. Most of these gene regulatory studies have consisted of largely unguided searches of genomic sequence for those with gene regulatory properties.

Computational sequence analysis provides three broadly different approaches for scanning genomic sequence to identify those regions predicted to participate in gene regulation. First, inter-species sequence comparisons have been used to identify non-coding sequences that have a reasonable likelihood of having gene regulatory properties[4–8]. This is possible because sequences that mediate gene expression tend to be conserved between species. Sequence analysis of co-regulated genes within a species is a second approach for predicting regulatory elements. This strategy is based on

Box 1 | A lexicon of *cis*-acting regulatory elements

Promoter
Sequence of DNA near the 5′ end of a gene that acts as a binding site for RNA polymerase and from which transcription is initiated.

Enhancer
Control element that elevates the levels of transcription from a promoter, independent of orientation or distance[67].

Locus control region (LCR)
Confers tissue-specific temporally regulated expression of linked genes. LCRs function independently of position, but they are copy number dependent and open the nucleosome structure so that other factors can bind. LCRs affect replication timing and origin usage[30,68,69].

Boundary element/insulator
DNA sequence that prevents the activation or inactivation of transcription by blocking the effects of surrounding chromatin[70,71].

Silencer
Control element that suppresses gene expression independent of orientation or distance[72].

Matrix attachment region (MAR)/scaffold attachment region
DNA sequence that binds the nuclear scaffold and can affect transcription. These elements probably form higher-order looped structures within chromosomes and influence gene expression by separating chromosomes into regulatory domains[73].

the fact that few transcription factors exert their activity exclusively on single genes; rather, most bind to conserved sites in several genes to coordinate their expression. Accordingly, genes are thought to be co-regulated because they respond to similar regulatory pathways owing to shared non-coding sequence motifs that direct the binding of specific sets of shared transcription factors[9–17]. The third approach for the identification of gene regulatory sequences involves generating and analysing databases of known transcription-factor-binding sites and characterizing promoter regions[18–29].

Accompanying the expansion of large data sets that have resulted from genomic sequencing and genome-wide expression profiling, are new computational strategies that have been developed to contribute to the creation of a vocabulary that describes gene regulatory instructions contained in genomic DNA. So far, most examples of genomic approaches that have been used to identify sequences that participate in gene regulation have involved model organisms with small but sequenced genomes. This is beginning to change with the increasing availability of genomic data sets for mammals. Accordingly, this review focuses on current and likely future strategies for integrating these data sets with regulatory sequence information.

Comparative genomic sequence analysis

The potential functions of conserved non-coding sequences are numerous, and include roles in chromosomal assembly and replication as well as gene regulation. Compelling support for the conservation of sequence-based regulatory information across species comes from a diverse set of experimental approaches. Most importantly, this support includes the DNA sequence conservation of experimentally defined regulatory elements among mammals[7,8,30]. This evidence is consistent with the numerous transgenic studies illustrating that genes from various mammals, when introduced into mice as genomic transgenes, are nearly always expressed in a manner that mimics their expression in the natural host[31]. Particularly telling results from transgenic studies include the rare situation in which the mouse lacks an orthologue of a transgene but nonetheless expresses it in a manner similar to the expression pattern of the gene in its natural host[32]. An example of this is the apolipoprotein (a) gene (*LPA*), which recently evolved in old world monkeys and accordingly is not found in rodents. When a large human genomic transgene (250 kb) containing *LPA* was introduced into the mouse genome, its tissue-specific expression, as well as aspects of its expression response to environmental factors, mimicked those found in humans. These studies are consistent with the existence of highly conserved instructions that are embedded in the non-coding sequence of mammalian genomes, which regulate the expression of neighbouring genes.

A key question that arises recurrently in contemplating a comparative genomic approach for identifying regulatory sequences is which species should be compared. The availability of genomic sequence for a wide variety of vertebrates would enable the reconstruction of mammalian evolution and would facilitate the determination of sequence features that are common or unique to each species. However, a plethora of completely sequenced mammalian genomes does not seem like an immediate prospect. The practical reality based on today's sequencing throughput and cost is that decisions will need to be made as to which organisms should be studied first to best annotate and understand the human sequence. From sequence comparisons that have been done, it is clear that no one species when compared to human will be completely informative for addressing all regulatory issues. For example, in addition to each species evolving independently since their last common ancestor, different regions of a genome seem to evolve at different rates. Some regions of the genome have evolved quickly, such as the β-globin locus control region (LCR)[7,8,33], allowing easy identification of conserved non-coding sequence among closely related mammals; other regions have evolved slowly, such as the T-cell receptor loci[34,35], and might require sequence comparisons between distantly related mammals or other vertebrates (for example, marsupials, birds, reptiles or fish). As no single organism can be used to annotate the entire human genome, several organisms should be used for comparative analysis depending on the human genomic interval and the question being investigated.

Human–mouse sequence comparison. As a prelude to large-scale sequence comparisons for identifying regulatory sequences, the analysis of the β-globin cluster has acted as a model for defining the properties of both local and distant non-coding regulatory elements. The various *cis*-regulatory elements, including the β-globin LCR, have been intensively studied and functionally characterized using various wet-laboratory approaches long before human and mouse genomic sequences of this

GEL SHIFT ASSAY
A gel-based assay in which proteins that bind to a DNA fragment are detected by virtue of the reduced migration of the DNA. The assay is often used to detect transcription factor binding.

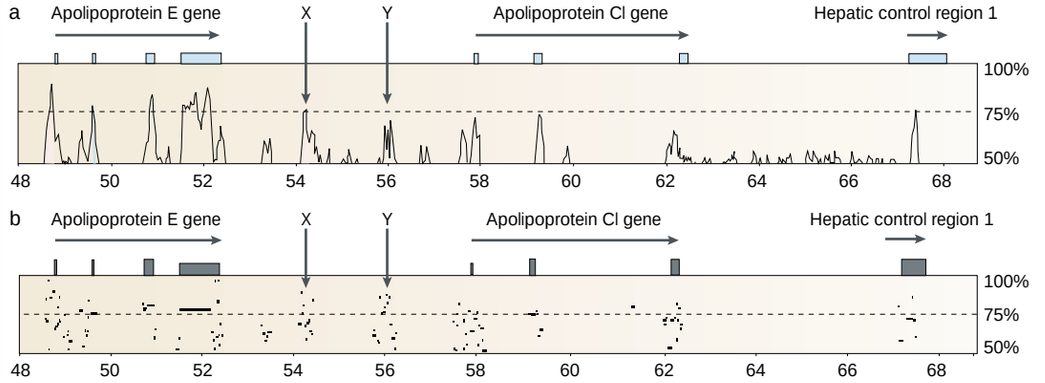
Box 2 | Computational approaches to cross-species sequence comparisons

One of the main effects of computational science on molecular biology has been the development of algorithms to detect conservation between sequences. Local alignment tools, such as BLAST[74], were primarily developed to rapidly identify sequence similarity between a relatively short query sequence and a large sequence database. By contrast, cross-species comparisons require accurate alignment of a small number of large contiguous sequences. Whereas local alignments have been used successfully for cross-species genomic sequence comparisons, global alignment algorithms provide an overall view that specifies how two large genomic sequences fit together. Once the pieces of a large genomic interval have been aligned, smaller regions of conservation in this interval can be identified[23,75,76].

Software programs have also been developed to visualize sequence alignment outputs. VISTA[77] (visualization tool for alignment) combines a global alignment program with a graphical tool for analysing alignments that allows the identification of conserved coding and non-coding sequences between species. The output creates a graphical plot (**a**) in which the horizontal axis represents the human genomic sequence and the vertical axis indicates the percentage of identical nucleotides in a predefined interval between human and another species across the alignment. The illustration shows an analysis using a 100-nucleotide window, which slides at 40-nucleotide increments. A similar program, PIPMaker[78], has also been used extensively in comparative analyses. After a local sequence alignment that uses a modified version of BLAST, a percentage identity plot (PIP) is generated (**b**). The PIP indicates regions of similarity based on the percentage identity of each gap-free segment of the alignment (the number of matches in the region divided by the length of the region). These percentages are shown using bars, with the height of each bar indicating the percentage identity in the corresponding gap-free segment. These, as well as other newly developed sequence visualization tools, allow investigators to analyse sequence data from two (or more) species to visually identify conserved non-coding regions in the vicinity of genes of interest.

In the illustrations, the orthologous apolipoprotein E gene (*APOE*) clusters in human and mouse are compared. Over the past two decades, an enormous amount of effort has been put into understanding the structure and regulation of the gene cluster on human chromosome 19 that contains *APOE*, because of its involvement in cardiovascular biology and Alzheimer disease. A significant finding was a pair of hepatic-specific enhancer elements (HCRs) located ~20 kb downstream of the human *APOE* gene.

VISTA and PipMaker readily identify the previously known HCR elements, based on high levels of sequence identity between human and mouse non-coding sequences. In addition to identifying previously known control elements in the *APOE* region, this analysis also identified several novel conserved non-coding sequence elements (X and Y). These conserved segments downstream of *APOE* represent potential regulatory sequences that warrant further experimental analysis.



region were available[36]. The β-globin LCR was identified and characterized initially through *in vitro* DNase I hypersensitivity assays and transgenic mouse studies[37]. After the experimental definition of the LCR, the availability of the genomic sequence of this region revealed that the LCR was highly conserved between human and mouse, in contrast to the surrounding non-coding sequences in this genomic interval[33]. This pattern of first identifying *cis*-regulatory sequences experimentally, and then learning that they are conserved at the sequence level among mammals, has been repeated many times[30]. With the recent availability of large amounts of genomic sequence from human and mouse, the chronology of experimental and comparative sequence studies is being reversed, with sequence analysis now preceding and directing ensuing experimental studies. Furthermore, several computational tools have been developed for the purpose of genomic sequence comparisons (BOX 2).

The largest human–mouse sequence comparison that aimed to identify regulatory elements was a study in which ~1 Mb of human chromosome region 5q31 (including 5 interleukins (IL) and 18 other genes) was compared to the orthologous region of mouse chromosome 11 (REF. 7). This cross-species annotation of sequence identified elements of 100 bp or greater that were over 70% conserved between human and mouse. These inclusion criteria were chosen on the basis of several studies in which long-range regulatory elements that showed experimentally determined functional properties (for example, β-globin LCR) were characterized and

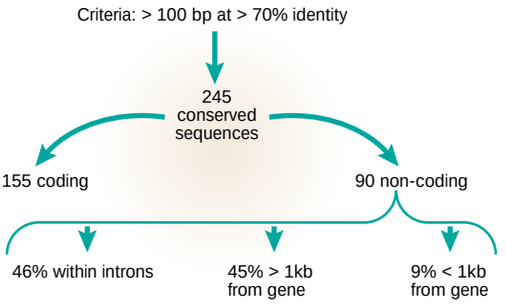


Figure 1 | **Classification of conserved human–mouse sequences.** In this 1-Mb survey of cross-species sequence comparison from a region of human chromosome 5 and mouse chromosome 11, conserved elements were defined as orthologous sequences greater than 100 bp with greater than 70% identity[7]. Of the 245 conserved elements that met these criteria, 155 (63%) were found in coding regions and 90 (37%) were found in non-coding regions. The 90 non-coding-conserved sequences could be classified further on the basis of whether they fell within an intron (46%), greater than 1 kb from a gene (45%) or less than 1 kb from a gene (9%).

found to meet these criteria. A survey of the characteristics and distribution of the 90 conserved sequences identified in this 1-Mb analysis is shown in FIG. 1. It was encouraging that experimentally characterized gene regulatory elements known to reside within this 1-Mb interval, such as the granulocyte-macrophage colony stimulating factor (*GMCSF*) enhancer, were readily identified by the human–mouse sequence comparison. Twelve out of fifteen conserved elements analysed by low stringency Southern analysis seem to be single copy, suggesting that a high percentage of these conserved elements are likely to be unique in the human genome.

After the computational analysis of the human–mouse 1-Mb region, the properties of a single conserved non-coding sequence, located in the 15-kb interval between *IL4* and *IL13*, were studied. Out of the 90 other conserved sequences that met similar criteria, this particular element was assigned priority for in-depth characterization for two reasons: the element was the largest conserved non-coding sequence identified in the 1-Mb interval (400 bp at ~87% identity between human and mouse); and previous studies indicated that *IL4* and *IL13* might be co-regulated in a subset of helper T cells, (T_H2 cells), raising the possibility that the element might participate in the regulation of both genes. To characterize the function of this sequence, a 450-kb yeast artificial chromosome (YAC) transgene (containing the putative human regulatory element flanked by *loxP* sites, and containing nine other human genes including those that encode the T_H2 specific cytokines IL4, IL5 and IL13) was studied in mice. The CRE RECOMBINASE SYSTEM was used to create separate lines of mice that contained the YAC transgene at a single integration site (both with and without the conserved element). Whereas the expression of six of the nine human genes contained on the human YAC were unaffected by the presence or absence of the conserved element, the absence of the element markedly altered the expression of the three T_H2-specific human cytokines. A marked reduction in the number of T_H2 cells expressing the human *IL4*, *IL5* and *IL13* genes, which are separated by more than 120 kb at the 5q31 locus, was noted in animals lacking the element compared with those with the element. These studies illustrate the complexity of long-range regulatory elements and the power of comparative biology to discover and to decipher the properties of such conserved regulatory elements.

Although these transgenic studies attribute a function to a single conserved non-coding sequence, the functional significance of the remaining 89 conserved sequences discovered in this 1-Mb analysis remains uncertain. The analysis of an orthologous 200-kb segment of the 1-Mb interval in the dog genome supports the view that most of the elements have been actively maintained[38]. In this analysis, approximately two-thirds of the conserved elements identified from any of the two-way comparisons (human–dog, human–mouse and dog–mouse) were shown to be present in all three species.

A second recent study, in which comparative sequence analysis identified gene regulatory sequences before functional studies, involved the analysis of a genomic interval that contains the stem cell leukaemia (*SCL*) gene[8]. In this study, investigators sequenced ~320 kb of genomic DNA that includes the *SCL* locus in human, mouse and chicken. Sequence comparisons showed the presence of numerous regions of homology within non-coding DNA. Comparisons between human and mouse sequence identified all previously defined *SCL* enhancers. In addition, one of the peaks of the human–mouse homology that did not correspond to a known enhancer was functionally characterized in a transgenic *Xenopus* reporter assay. This approach identified a novel neuronal enhancer that lies adjacent to the *SCL* gene. An interesting point in this study was the use of orthologous sequence from an evolutionarily distant species — the chicken. Although the chicken sequence contained evolutionary conservation of a subset of the conserved elements identified by human–mouse comparisons, several human–mouse conserved elements that were experimentally verified to be functional were not conserved in the chicken. This suggests that although orthologous chicken sequence can be useful to assign priority to a subset of sequence elements for further experimental study, this screening strategy alone might miss many mammalian regulatory sequences.

Caveats. In the comparative studies involving human 5q31, the *SCL* locus and the β-globin locus, both non-coding regulatory elements that had previously been identified experimentally, and elements identified initially from sequence-based analyses, stood out as islands of conserved sequence in seas of less conserved non-coding sequence. These results indicate that, at least in these regions of the human genome, the evolutionary distance between human and mouse is adequate to facilitate the identification of regulatory elements. However, various caveats need to be considered before using this approach to identify regulatory elements, including the finding that some regions of the genome are highly conserved between human and mouse both

CRE RECOMBINASE SYSTEM
A method in which the Cre enzyme catalyses recombination between *loxP* sequences. If the *loxP* sequences are arranged as a direct repeat, recombination will delete the DNA between the sites.

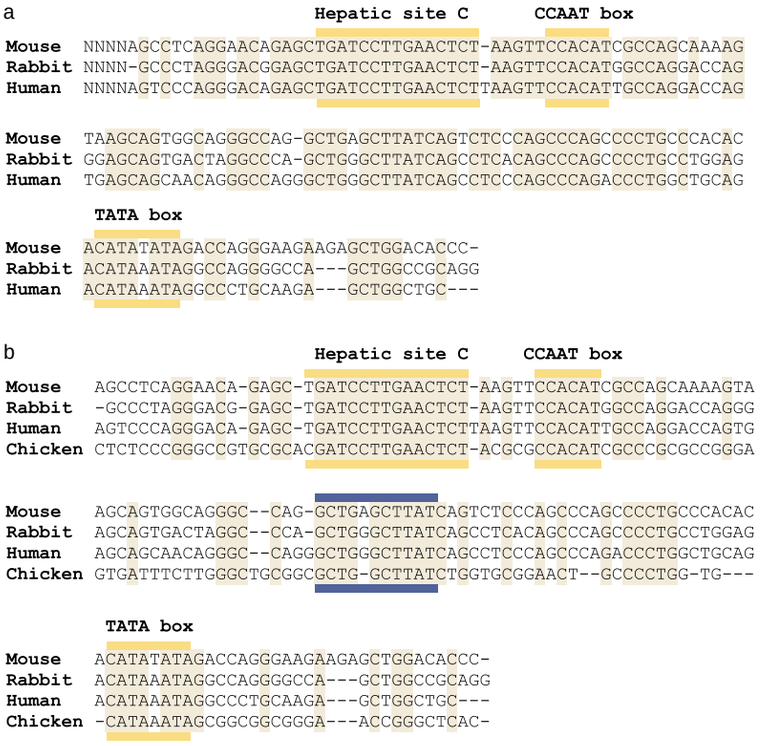


Figure 2 | **Identifying transcription-factor-binding sites.** To illustrate the power of using multi-species comparative genomic sequence analysis to identify transcription-factor-binding sites, the upstream region of the well-studied apolipoprotein AI (*ApoAI*) gene was examined. **a** | Roughly 150 bp upstream of the predicted *ApoAI* transcription start site in human, mouse and rabbit was compared. This comparison indicates high levels of sequence conservation across the entire region in these mammals, making it difficult to assign priority to any sequences that were more likely to be transcription-factor-binding sites. **b** | To identify regulatory sequences more precisely, the orthologous region of the chicken *ApoAI* gene was added. This decreased the level of conservation greatly. Importantly, the highest levels of conservation were found in regions previously shown to be important in gene regulation (yellow). Both the CCAAT box and the TATA box, important in core promoter activity, are almost perfectly conserved across all four species. In addition, hepatic enhancer site C, experimentally shown to be necessary for *ApoAI* liver expression, reveals strong sequence conservation (14 of 15 bp are conserved across all 4 species). The other novel conserved block (blue) that was revealed by comparative analysis has yet to be assigned a biological function.

in coding and non-coding regions. This is exemplified by the analysis of 100 kb of contiguous DNA in the Cδ and Cα regions of the α/δ T-cell receptor loci of human and mouse[34,35]. The entire region in this study averaged 71% similarity, with non-coding sequences being conserved only slightly less than coding sequences. No islands of discernibly increased conservation in non-coding regions were identified.

Another reason why comparative sequence analysis cannot be used as an exclusive approach for identifying regulatory elements is that several experimentally characterized regulatory elements have failed to show inter-species sequence conservation. An example of this limitation is provided by the α-like globin gene cluster, which has been compared in human and rabbit, species that are more closely related than human and mouse[39]. Experimentally defined regulatory elements lacked significant sequence homology between these species. So, although comparative studies are likely to identify a large number of functionally important sequences on the basis of a high degree of conservation, it is important to point out that a fraction of gene regulatory elements will be missed regardless of which species are compared.

The inter-species sequence comparison approach can be rapidly applied to large stretches of genomic sequence as more sequence becomes available. This will clearly help to decide which regions of non-coding DNA are most likely to participate in gene regulation. The challenge after high-throughput sequence analyses will be to determine the function of the identified conserved sequence. Complementary high-throughput approaches to add functional information to putative regulatory sequences are desperately needed. Although still in its infancy, one such strategy that is being developed is to crosslink transcription factors to the DNA sequences to which they bind *in vivo*[40]. Importantly, this type of approach has the potential to be practised on a genome-wide scale and would offer an independent method to determine which sequences are truly recognized by proteins involved in gene regulation. In addition, the integration of transcription-factor-binding site prediction programs with multi-species genome comparisons is another means for providing a high-throughput computational priority assignment of conserved sequences that are likely to be involved in gene regulation (FIG. 2).

Looking for known DNA sequence motifs

Comparative sequence-based approaches are well suited to identify large conserved non-coding sequences (such as LCRs) that are likely to be composed of multiple regulatory elements. To identify specific regulatory binding sites, biochemical methods and computational strategies that do not require cross-species sequence comparisons have been exploited. For instance, the presence of CPG ISLANDS serves as one means to identify regulatory sequences — in this case, putative promoter-containing regions. In addition, genetic and biochemical analyses have been intensively carried out on a gene-by-gene basis to identify the precise sequences to which many transcription factors bind.

The availability of consensus target sequences for many of the known transcription factors has been used to construct databases that can be searched to identify potential transcription-factor-binding sites in a DNA sequence. Although useful data sets have been generated, the identification of transcription-factor-binding sites still presents a formidable challenge. Despite the fact that some transcription factors bind to highly specific DNA sequences, most have a small invariant core sequence (about 4–6 bp) surrounded by a variable number of degenerate nucleotides. Several strategies are used to classify transcription-factor-binding sites in databases. The most inflexible is to use a single unambiguous sequence to categorize a specific binding site (for example, TATAA). Alternatively, consensus sequences can incorporate ambiguous positions (for example, TARAA, where R=A or G)[18,41]. Most recently, position-weighted matrices have been used to estimate the likelihood that a given sequence

CPG ISLANDS
Sequences of at least 200 bp with greater than 50% G+C content and high CpG frequency.

Field	Description	
AC	M00134	**(Accession number)**
ID	V$HNF4	**(Identifier)**
DT	22.05.1995	**(Date created)**
DT	18.10.1995	**(Date updated)**
NA	HNF-4	**(Name of the binding factor)**
DE	hepatic nuclear factor 4	**(Short factor description)**

Number of sequences with a given nucleotide at that position

Nucleotide position	A	C	G	T	Consensus
P0	A	C	G	T	
01	10	4	4	6	N
02	6	9	7	5	N
03	12	6	7	6	N
04	12	3	14	3	R
05	2	0	29	1	G
06	5	2	17	8	G
07	3	8	10	11	N
08	1	23	1	7	**C**
09	27	1	3	1	**A**
10	29	0	3	0	**A**
11	26	0	5	1	**A**
12	3	0	28	1	**G**
13	3	1	16	12	K
14	2	6	6	18	T
15	0	24	1	7	C
16	22	4	4	2	A
17	9	9	6	6	N
18	7	5	13	5	N
19	8	3	6	7	N

BA 32 binding sites from 24 genes **(Statistical basis)**
CC compiled sequences **(Comments)**

Figure 3 | **TRANSFAC and position-weighted matrices.** The hepatic nuclear factor 4 (HNF4) position-weighted matrix was obtained from TRANSFAC to illustrate the features of these data files[28]. The various descriptions contained in this entry are indicated in parentheses. For instance, under the BA field, it is indicated that the matrix was compiled from sequences of 32 HNF4-binding sites from 24 genes. Within the matrix, the number of sequences that contain a given nucleotide is indicated, for each position near the experimentally determined binding site. Consensus sequences for the binding site are generated on the basis of the frequency of a given nucleotide at that position. The core consensus binding sequence is shown in bold (CAAAG).

binds to a specific transcription factor[19]. In this method, experimental data are used to assign a score for each base at each position in the transcription-factor-binding site. The score is based on a set of known binding sites and the frequency with which each base is found at each position (FIG. 3).

At present, the most widely used transcription factor database is TRANSFAC[28], which was introduced about a decade ago to catalogue experimentally derived data on transcription factors and their binding sites. It began as a computer-readable FLAT FILE, and has subsequently evolved into a RELATIONAL DATABASE. Recently, TRANSFAC has been linked with other regulatory element databases, including the Transcription Regulatory Region Database (TRRD) and COMPEL, a database of composite regulatory elements[21,22]. TRANSFAC is continually updated as more experimental data become available and is used by a large number of sequence analysis programs to identify potential binding sites[24,25,27,29,42].

FLAT FILE
A computer readable file or database in which records are not connected or 'related'. Similar to a card index.

RELATIONAL DATABASE
A storage format in which data items can be stored in separate files but linked together to form different relations. This system allows greater flexibility than a flat file format.

One main difficulty with the output from transcription factor database searches is the large number of false-positive returns. The short length and degenerate nature of transcription-factor-binding sites account for most of these misleading predictions. For instance, the unambiguous sequence TATAA is expected once every 1,024 bp by chance, which predicts 30 million potential binding sites in a mammalian genome. Therefore, the vast majority of predicted binding sites in mammalian genomic sequence are biologically non-functional. Various strategies have been used to sift the output of a transcription factor database search to decrease the number of false-positive returns. Power can be gained by taking advantage of the sequence context in which a predicted binding site is found. In the TATAA example, higher statistical scores can be assigned if the site is found within 25–30 bp of a predicted transcription initiation sequence. Predictions can be further strengthened if a transcription factor is known to function as a dimer, and two similar adjacent binding sites are found. Other approaches to reduce false-positive predictions include detecting clustered or composite binding sites[43–45]. Despite these various ways to minimize the number of false-positive binding sites, even those sequences that meet the most stringent criteria might still be non-functional in a genomic context. For instance, the binding site might be inaccessible owing to chromatin structure or blockage by other proteins.

An equally important problem is the large number of binding sites that are missed in such a transcription factor database search (false-negatives). Only a fraction of mammalian transcription factors and their binding sites are known and available for comparison, leaving a large set of unknown transcription-factor-binding sites. Comparative sequence information can help by signalling the presence of novel binding sites that might not have been predicted using sequence from a single species. Binding sites found in human sequence that are also found in orthologous mouse and other mammalian sequences are far more likely to be real than those found only in humans (FIG. 2). The term 'phylogenetic footprint' has been used to refer to these short orthologous sequences that are conserved over 6 bp or more[46–48].

Another strategy to reduce the number of false-positive predictions in regulatory sequence identification is to define the specific category of non-coding genomic sequence included in the analyses. For example, the Eukaryotic Promoter Database (EPD) is an annotated non-redundant collection of eukaryotic RNA polymerase II promoters for which the transcription start site has been determined experimentally[26]. This database is confined to sequences that are found immediately upstream of the transcription start site and is therefore limited in its ability to identify distant regulatory elements. However, the EPD does have a significant value in comparative genomic analysis of sequences that lie adjacent to orthologous genes in several species, or for intra-species comparison of co-expressed genes. As for the strategies discussed above, the identification of conserved sequences upstream of transcription start sites across several species facilitates the identification of short-range transcription-factor-binding sites (FIG. 2).

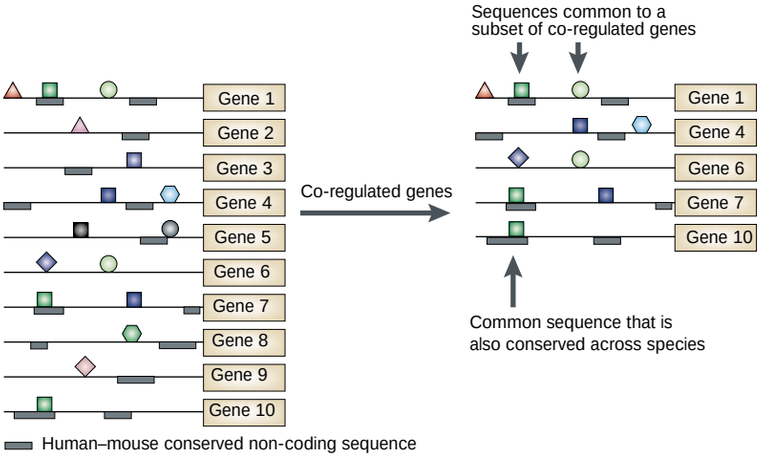


Figure 4 | **Combining expression data and sequence conservation.** This illustration represents a hypothetical set of genes with various sequence motifs (coloured symbols) upstream of their transcription start sites. Co-regulated genes are identified by transcriptional profiling. Examination of motifs among these co-regulated genes identifies common motifs (green). In this example, conserved non-coding sequences have also been identified in the co-regulated genes, and only one of the common motifs is conserved. This conserved element now functions as a strong candidate sequence participating in the coordinated regulation of these genes.

In addition to identifying regulatory sequences on the basis of similarity to known transcription-factor-binding sites, sequence-based strategies that search for more general features of DNA regions associated with gene regulation also exist. It has been shown that ~50% of mammalian promoters are associated with one or more CpG islands, although the presence of CpG islands is not always indicative of a promoter[49]. Biochemical approaches to identify or map CpG islands involved cleavage with restriction enzymes that preferentially cut CpG-rich DNA[50–54]. More recently, sequence analysis software has been developed to search for CpG-rich DNA. This included the analysis of the complete sequence of human chromosome 22 (REF. 55). In both the computational and biochemical approaches for identifying CpG islands, a significant number of elements are identified that have no promoter activity. To overcome this weakness, a new screening strategy has recently been developed by Ioshikhes and Zhang to reduce false-positive predictions while maintaining high sensitivity[56]. By characterizing the sequence composition of CpG islands that are associated or are not associated with gene promoters, they showed that CpG islands in close proximity to promoters have increased length, G+C content, and CpG dinucleotide frequency. Their refined search based on these criteria resulted in identifying more than 93% of a known set of promoters that contain CpG islands. The application of this and other computational strategies for identifying sequences with regulatory characteristics should facilitate the large-scale identification of likely promoter elements throughout a mammalian genome.

*MLU*I CELL-CYCLE BOX
An 8-bp motif (ACGCGTNA) that promotes the transcription of genes involved in DNA replication in yeast.

Expression profiling

A comprehensive understanding of the gene regulation of an organism at the genome-wide level requires the identification of those sequences likely to be bound by factors that affect the synthesis of RNA. If one considers the limited number of transcription factors compared with the total number of genes in a genome, it is clear that few transcription factors or other DNA-binding proteins bind exclusively to the non-coding sequence of a single gene. Co-expression of genes therefore reflects regulation by common transcription factors, and intra-species sequence comparisons of non-coding DNA near co-expressed genes are likely to be a fruitful strategy for identifying common sequences that are important in coordinated gene regulation. This approach is becoming increasingly feasible with the availability and coupling of genomic sequence and genome-wide expression profiling data. Because this strategy has been used most extensively for organisms with smaller genomes, such as yeast, we describe those studies first and then consider whether the same approach can be applied successfully to mammals[10,12,14–17,57–59] (FIG. 4).

Transcriptional profiling in yeast. In one of the first studies of its kind, Chu *et al.* examined the transcription profiles of yeast during sporulation by microarray analysis of the ~6,000 predicted yeast open reading frames[9]. Cluster analysis, aimed at identifying potentially co-regulated genes, yielded genes with similar expression profiles corresponding to early, middle, middle-late and late events during sporulation. This analysis identified 62 genes that were induced within 30 minutes after transfer to sporulation media. Of these early induced genes, roughly one-third had consensus URS1 transcription-factor-binding sites within 600-bp upstream of their start codons. Some of the binding sites had previously been shown to be functional. Although this represents only a single example of the data generated in this study, in each of the various stages of sporulation many of the genes that showed similar expression profiles were enriched for various transcription-factor-binding sites.

In a more recent parallel study, Tavazoie *et al.* clustered yeast genes on the basis of their expression profile during the mitotic cell cycle[11]. Genes that showed similar expression patterns were then examined for sequence similarities within 600-bp upstream of the start codon of each gene. One finding from this analysis was that more than 50% of a clustered set of genes, for which expression peaks in G1 phase, contained a predicted *MLU*I CELL-CYCLE BOX. In addition, many other clusters of co-expressed genes were significantly enriched for certain binding sites that had been previously determined to regulate individual members of a cluster. These, as well as other studies in yeast, support the hypothesis that identifying subsets of genes that are co-expressed can assign priority to genes for intra-sequence comparisons and facilitate the discovery of common sequence binding sites that are likely to be regulated by similar factors.

Mammalian intra-species analysis. The absence of the complete genomic sequence for human and mouse is a deficiency that will soon be corrected.

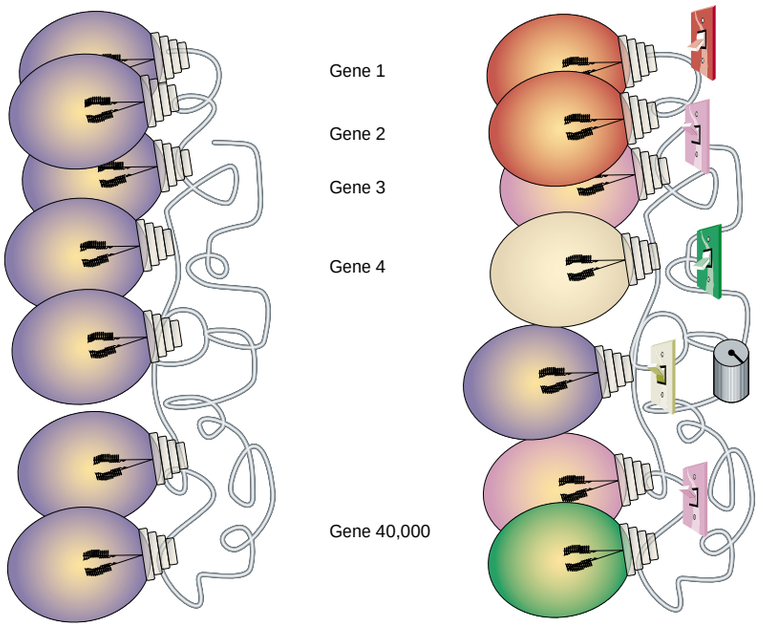


Figure 5 | **Annotating the genome.** Depiction of two of the stages in annotating the human genome. **a** | Current successes in large-scale sequencing and gene identification have provided the identity and physical location of a significant fraction of all human genes (light bulbs). **b** | The future development and implementation of regulatory sequence identification strategies will notably increase our understanding of chromosomal structure, regulatory elements (switches and rheostats) and expression patterns such as co-regulation (light bulbs of similar colour).

However, using intra-species sequence comparisons to identify regulatory sequences in mammals presents challenges that do not exist in simple eukaryotes such as yeast. For instance, yeast regulatory elements are nearly always found within several hundred base pairs of the translation start site. By contrast, mammalian regulatory elements are frequently found much farther away. Another challenge is the enormous size and complexity of mammalian genomes. In addition to being ~200 times larger than the yeast genome, more than 95% of the mammalian genome comprises non-coding sequences, compared with less than 50% in yeast. Unlike sequence comparison of co-regulated genes in yeast, for which there is a relatively small spatial window for identifying conserved sequences, intra-species comparisons in mammals require the analysis of large intervals of sequence surrounding co-regulated genes. Clearly, this large increase in complexity increases the chance for random sequence alignments.

Because mammalian genomes are far too complex, even in the general proximity of co-regulated genes, to look for sequences that are overrepresented in all non-coding DNA, additional strategies to assign priority to those non-coding sequences likely to encode regulatory information are required. One way to assign priority to non-coding sequences that are suspected to be involved in coordinate gene regulation is to couple this analysis to data derived from cross-species sequence comparisons. In a recent study, Wasserman *et al.* did human–rodent comparative sequence analyses upstream of a set of 28 orthologous genes with upregulated expression in skeletal muscle[13,60]. These 28 genes contain 75 sequence-specific binding sites previously shown to be functional. Using a modified algorithm for sequence comparison, 19% of non-coding sequences were conserved between human and rodents in the analysed genomic interval, including 74 of the 75 experimentally defined muscle-tissue transcription-factor-binding sites. In addition, novel muscle-specific binding sites were predicted in conserved regions, warranting further experimental investigation. As illustrated here, inter-species sequence analyses can assist in the identification of those sequences involved in coordinate gene regulation. The complexity of mammalian genomes will certainly require that these and other strategies are applied, to identify regulatory elements in sequences close to coordinately regulated genes.

Regulatory frontiers

We have entered an era of enormous increases in the availability of genomic data sets from a wide variety of animals. Inter- and intra-species sequence analyses, coupled with the development of algorithms to search genomic databases, provide important tools for the identification of gene regulatory elements at a scale not previously possible. As more animal genomes are sequenced, deeper sequence alignments will contribute further to the definition of putative regulatory elements and to the determination of the evolutionary extent of regulatory sequence conservation across species.

The application of comparative genomics to study gene regulation has focused largely on the identification of shared regulatory sequences to explain similar patterns of gene expression between species. By contrast, the differences in gene regulation between organisms, and the role of these differences in speciation, are fascinating issues that have only just begun to be examined. These topics are poised to be explored in the future with the availability of genomic data from several species that were separated at different times in evolution. For example, sequence conservation unique to the class Mammalia and not found in Aves and Reptilia might help identify the genetic causes of the biochemical and structural differences that make mammals unique. The fact that some species can differ enormously phenotypically, despite having been derived relatively recently from a common ancestor, suggests that different ways of regulating the expression of a fixed set of shared genes probably contributes to these differences[61]. A prime example of this is the whale and the hippopotamus, which shared a common founder only 20–40 million years ago[62–66]. Comparative analyses of different expression patterns between multiple species offer an important window into the molecular basis of phenotypic differences between species.

In the near future we are likely to know the linear sequence of the human genome as well as the precise loca-

tion of its complete set of genes (FIG. 5). Current efforts to develop a vocabulary for regulatory sequences should facilitate the large-scale identification of non-coding gene regulatory switches and an understanding of how they control the expression of a significant fraction of those genes. These attainable goals will contribute to marked increases in our understanding of mammalian biology.

Links

DATABASE LINKS *LPA* | *GMCSF* | *IL4* | *IL13* | IL5 | *SCL*
FURTHER INFORMATION TRANSFAC | Transcription Regulatory Region Database | COMPEL | Eukaryotic Promoter Database | VISTA | PipMaker | hepatic nuclear factor 4 (HNF4) position-weighted matrix

1. Durick, K., Mendlein, J. & Xanthopoulos, K. G. Hunting with traps: genome-wide strategies for gene discovery and functional analysis. *Genome Res.* **9**, 1019–1025 (1999).
2. Fukushige, S. & Ikeda, J. E. Trapping of mammalian promoters by Cre-lox site-specific recombination. *DNA Res.* **3**, 73–80 (1996).
3. Asoh, S., Lee-Kwon, W., Mouradian, M. M. & Nirenberg, M. Selection of DNA clones with enhancer sequences. *Proc. Natl Acad. Sci. USA* **91**, 6982–6986 (1994).
4. Duret, L. & Bucher, P. Searching for regulatory elements in human noncoding sequences. *Curr. Opin. Struct. Biol.* **7**, 399–406 (1997).
5. Hardison, R. C., Oeltjen, J. & Miller, W. Long human–mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Res.* **7**, 959–966 (1997).
6. Hardison, R. C. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.* **16**, 369–372 (2000).
   An excellent review of comparative sequence analyses, limitations and successes.
7. Loots, G. G. *et al.* Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288**, 136–140 (2000).
8. Gottgens, B. *et al.* Analysis of vertebrate SCL loci identifies conserved enhancers. *Nature Biotechnol.* **18**, 181–186 (2000).
   References 7 and 8 are early examples of the use of human–mouse comparative sequence analyses for assigning priority to regions of DNA to screen for functional properties.
9. Chu, S. *et al.* The transcriptional program of sporulation in budding yeast. *Science* **282**, 699–705 (1998); erratum **282**, 1421 (1998)
10. Spellman, P. T. *et al.* Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* **9**, 3273–3297 (1998).
11. Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. & Church, G. M. Systematic determination of genetic network architecture. *Nature Genet.* **22**, 281–285 (1999).
    References 9, 10 and 11 provide excellent examples of yeast microarray data and how they can be used to cluster pathway-related genes on the basis of similar expression patterns.
12. Zhu, J. & Zhang, M. Q. Cluster, function and promoter: analysis of yeast expression array. *Pac. Symp. Biocomput.* 479–490 (2000).
13. Wasserman, W. W. & Fickett, J. W. Identification of regulatory regions which confer muscle-specific gene expression. *J. Mol. Biol.* **278**, 167–181 (1998).
14. Niehrs, C. & Pollet, N. Synexpression groups in eukaryotes. *Nature* **402**, 483–487 (1999).
15. Lockhart, D. J. & Winzeler, E. A. Genomics, gene expression and DNA arrays. *Nature* **405**, 827–836 (2000).
    A significant review of the numerous applications of using DNA arrays to understand biological processes.
16. Hughes, J. D., Estep, P. W., Tavazoie, S. & Church, G. M. Computational identification of *cis*-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *J. Mol. Biol.* **296**, 1205–1214 (2000).
17. Zhang, M. Q. Promoter analysis of co-regulated genes in the yeast genome. *Comput. Chem.* **23**, 233–250 (1999).
18. Faisst, S. & Meyer, S. Compilation of vertebrate-encoded transcription factors. *Nucleic Acids Res.* **20**, 3–26 (1992).
19. Frech, K., Herrmann, G. & Werner, T. Computer-assisted prediction, classification, and delimitation of protein binding sites in nucleic acids. *Nucleic Acids Res.* **21**, 1655–1664 (1993).
20. Ghosh, D. Object-oriented transcription factors database (ooTFD). *Nucleic Acids Res.* **28**, 308–310 (2000).
21. Heinemeyer, T. *et al.* Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. *Nucleic Acids Res.* **26**, 362–367 (1998).
22. Kel-Margoulis, O. V., Romashchenko, A. G., Kolchanov, N. A., Wingender, E. & Kel, A. E. COMPEL: a database on composite regulatory elements providing combinatorial transcriptional regulation. *Nucleic Acids Res.* **28**, 311–315 (2000).
23. Morgenstern, B. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics* **15**, 211–218 (1999).
24. Prestridge, D. S. SIGNAL SCAN 4.0: additional databases and sequence formats. *Comput. Appl. Biosci.* **12**, 157–160 (1996).
25. Prestridge, D. S. Computer software for eukaryotic promoter analysis. *Methods Mol. Biol.* **130**, 265–295 (2000).
26. Perier, R. C., Praz, V., Junier, T., Bonnard, C. & Bucher, P. The eukaryotic promoter database (EPD). *Nucleic Acids Res.* **28**, 302–303 (2000).
27. Quandt, K., Frech, K., Karas, H., Wingender, E. & Werner, T. MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.* **23**, 4878–4884 (1995).
28. Wingender, E. *et al.* TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.* **28**, 316–319 (2000).
29. Werner, T. Computer-assisted analysis of transcription control regions. Matinspector and other programs. *Methods Mol. Biol.* **132**, 337–349 (2000).
30. Li, Q., Harju, S. & Peterson, K. R. Locus control regions: coming of age at a decade plus. *Trends Genet.* **15**, 403–408 (1999).
    A detailed summary of our current understanding of the β-globin locus control region.
31. Lacy, D. A. *et al.* Faithful expression of the human 5q31 cytokine cluster in transgenic mice. *J. Immunol.* **164**, 4569–4574 (2000).
32. Frazer, K. A., Narla, G., Zhang, J. L. & Rubin, E. M. The apolipoprotein(a) gene is regulated by sex hormones and acute-phase inducers in YAC transgenic mice. *Nature Genet.* **9**, 424–431 (1995).
    A transgenic study supporting the commonality of gene regulation between species.
33. Jimenez, G., Gale, K. B. & Enver, T. The mouse β-globin locus control region: hypersensitive sites 3 and 4. *Nucleic Acids Res.* **20**, 5797–5803 (1992).
34. Hood, L., Rowen, L. & Koop, B. F. Human and mouse T-cell receptor loci: genomics, evolution, diversity, and serendipity. *Ann. NY Acad. Sci.* **758**, 390–412 (1995).
35. Koop, B. F. & Hood, L. Striking sequence similarity over almost 100 kilobases of human and mouse T-cell receptor DNA. *Nature Genet.* **7**, 48–53 (1994).
    An example of a large genomic region in human and mouse that is highly conserved, thus limiting regulatory sequence identification.
36. Ho, P. J. & Thein, S. L. Gene regulation and deregulation: a β-globin perspective. *Blood Rev.* **14**, 78–93 (2000).
37. Talbot, D. *et al.* A dominant control region from the human β-globin locus conferring integration site-independent gene expression. *Nature* **338**, 352–355 (1989).
38. Dubchak, I. *et al.* Active conservation of noncoding sequences revealed by three-way species comparisons. *Genome Res.* **10**, 1304–1306 (2000).
39. Hardison, R. *et al.* Sequence and comparative analysis of the rabbit α-like globin gene cluster reveals a rapid mode of evolution in a G+C-rich region of mammalian genomes. *J. Mol. Biol.* **222**, 233–249 (1991).
40. Bulyk, M. L., Gentalen, E., Lockhart, D. J. & Church, G. M. Quantifying DNA-protein interactions by double-stranded DNA arrays. *Nature Biotechnol.* **17**, 573–577 (1999).
41. Cavener, D. R. Comparison of the consensus sequence flanking translational start sites in *Drosophila* and vertebrates. *Nucleic Acids Res.* **15**, 1353–1361 (1987).
42. Werner, T. Models for prediction and recognition of eukaryotic promoters. *Mamm. Genome* **10**, 168–175 (1999).
43. Wagner, A. A computational genomics approach to the identification of gene networks. *Nucleic Acids Res.* **25**, 3594–3604 (1997).
44. van Helden, J., Andre, B. & Collado-Vides, J. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.* **281**, 827–842 (1998).
45. Wagner, A. Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes. *Bioinformatics* **15**, 776–784 (1999).
46. Tagle, D. A. *et al.* Embryonic ε- and γ-globin genes of a prosimian primate (*Galago crassicaudatus*). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J. Mol. Biol.* **203**, 439–455 (1988).
47. Vuillaumier, S. *et al.* Cross-species characterization of the promoter region of the cystic fibrosis transmembrane conductance regulator gene reveals multiple levels of regulation. *Biochem J.* **327**, 651–662 (1997).
48. Gumucio, D. L. *et al.* Evolutionary strategies for the elucidation of *cis*- and *trans*-factors that regulate the developmental switching programs of the β-like globin genes. *Mol. Phylogenet. Evol.* **5**, 18–32 (1996).
    References 46 and 48 illustrate the power of comparative genomic analyses through phylogenetic footprints of globin genes.
49. Antequera, F. & Bird, A. Number of CpG islands and genes in human and mouse. *Proc. Natl Acad. Sci. USA* **90**, 11995–11999 (1993).
50. Cross, S. H., Clark, V. H. & Bird, A. P. Isolation of CpG islands from large genomic clones. *Nucleic Acids Res.* **27**, 2099–2107 (1999).
51. John, R. M., Robbins, C. A. & Myers, R. M. Identification of genes within CpG-enriched DNA from human chromosome 4p16.3. *Hum. Mol. Genet.* **3**, 1611–1616 (1994).
52. Watanabe, T. *et al.* Isolation of estrogen-responsive genes with a CpG island library. *Mol. Cell. Biol.* **18**, 442–449 (1998).
53. Larsen, F., Gundersen, G. & Prydz, H. Choice of enzymes for mapping based on CpG islands in the human genome. *Genet. Anal. Tech. Appl.* **9**, 80–85 (1992).
54. Kato, R. & Sasaki, H. Quick identification and localization of CpG islands in large genomic fragments by partial digestion with HpaII and HhaI. *DNA Res.* **5**, 287–295 (1998).
55. Dunham, I. *et al.* The DNA sequence of human chromosome 22. *Nature* **402**, 489–495 (1999); erratum **404**, 904 (2000).
56. Ioshikhes, I. P. & Zhang, M. Q. Large-scale human promoter mapping using CpG islands. *Nature Genet.* **26**, 61–63 (2000).
57. Bucher, P. Regulatory elements and expression profiles. *Curr. Opin. Struct. Biol.* **9**, 400–407 (1999).
58. Greenfield, A. Applications of DNA microarrays to the transcriptional analysis of mammalian genomes. *Mamm. Genome* **11**, 609–613 (2000).
59. Hill, A. A., Hunter, C. P., Tsung, B. T., Tucker-Kellogg, G. & Brown, E. L. Genomic analysis of gene expression in *C. elegans*. *Science* **290**, 809–812 (2000).
60. Wasserman, W. W., Palumbo, M., Thompson, W., Fickett, J. W. & Lawrence, C. E. Human-mouse genome comparisons to locate regulatory sites. *Nature Genet.* **26**, 225–228 (2000).
61. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188**, 107–116 (1975).
    A landmark paper highlighting the large amount of sequence conservation between humans and chimpanzees, indicating that regulatory differences might account for the varying phenotypes between the two species.
62. Luo, Z. In search of the whales' sisters. *Nature* **404**, 235–237 (2000).
63. Arnason, U., Gullberg, A., Gretarsdottir, S., Ursing, B. & Janke, A. The mitochondrial genome of the sperm whale and a new molecular reference for estimating eutherian divergence dates. *J. Mol. Evol.* **50**, 569–578 (2000).
64. Ursing, B. M. & Arnason, U. Analyses of mitochondrial genomes strongly support a hippopotamus-whale clade. *Proc. R. Soc. Lond. B* **265**, 2251–2255 (1998).
65. Shimamura, M. *et al.* Molecular evidence from retroposons that whales form a clade within even-toed

ungulates. *Nature* **388**, 666–670 (1997).
66. Nikaido, M., Rooney, A. P. & Okada, N. Phylogenetic relationships among cetartiodactyls based on insertions of short and long interpersed elements: hippopotamuses are the closest extant relatives of whales. *Proc. Natl Acad. Sci. USA* **96**, 10261–10266 (1999).
67. Blackwood, E. M. & Kadonaga, J. T. Going the distance: a current view of enhancer action. *Science* **281**, 61–63 (1998).
68. Fraser, P. & Grosveld, F. Locus control regions, chromatin activation and transcription. *Curr. Opin. Cell Biol.* **10**, 361–365 (1998).
69. Grosveld, F. Activation by locus control regions? *Curr. Opin. Genet. Dev.* **9**, 152–157 (1999).
70. Bell, A. C. & Felsenfeld, G. Stopped at the border: boundaries and insulators. *Curr. Opin. Genet. Dev.* **9**, 191–198 (1999).
71. Geyer, P. K. The role of insulator elements in defining domains of gene expression. *Curr. Opin. Genet. Dev.* **7**, 242–248 (1997).
72. Ogbourne, S. & Antalis, T. M. Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. *Biochem J.* **331**, 1–14 (1998).
73. Hart, C. M. & Laemmli, U. K. Facilitation of chromatin dynamics by SARs. *Curr. Opin. Genet. Dev.* **8**, 519–525 (1998).
74. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
75. Batzoglou, S., Pachter, L., Mesirov, J. P., Berger, B. & Lander, E. S. Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res.* **10**, 950–958 (2000).
76. Delcher, A. L. *et al.* Alignment of whole genomes. *Nucleic Acids Res.* **27**, 2369–2376 (1999).
77. Mayor, C. *et al.* VISTA: Visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* (in the press).
78. Schwartz, S. *et al.* PipMaker — a web server for aligning two genomic DNA sequences. *Genome Res.* **10**, 577–586 (2000).

Acknowledgements

This research was supported by a Programs for Genomic Applications grant awarded to E.M.R. from the NHLBI and conducted at the E.O. Lawrence Berkeley National Laboratory, University of California, sponsored by the Department of Energy, as well as an appointment to the Alexander Hollaender Distinguished Postdoctoral Fellowship Program sponsored by the US Department of Energy, Office of Biological and Environmental Research, and administered by the Oak Ridge Institute for Science and Education (L.A.P.). We thank M. Biggin, J. Bristow, I. Dubchak, C. Prangeand D. Symula for their thoughtful comments.