Bioinformatics for the 'bench biologist': how to find regulatory regions in genomic DNA

Julie Nardone, Dong U Lee, K Mark Ansel & Anjana Rao

The combination of bioinformatic and biological approaches constitutes a powerful method for identifying gene regulatory elements. High-quality genome sequences are available in public databases for several vertebrate species. Comparative cross-species sequence analysis of these genomes shows considerable conservation of noncoding sequences in DNA. Biological analyses show that an unexpectedly high number of the conserved sequences correspond to functional *cis*-regulatory regions that influence gene transcription. Because research biologists are often unfamiliar with the bioinformatic resources at their disposal, this commentary discusses how to integrate biological and bioinformatic methods in the discovery of gene regulatory regions and includes a tutorial on widely available comparative genomics programs.

Whether primates are compared with fish or flies to nematodes, gene number does not correlate with differences in organism complexity[1]. Instead, complexity is conferred by variations in timing, abundance and localization of gene expression; differences in alternative splicing, transport and stability of mRNAs; and altered localization and post-translational modification of proteins. In particular, the evolution of diverse modes of transcriptional regulation has helped compensate for the 'frugal' supply of genes[2]. Transcription is initiated at gene promoters, but many classes of transcriptional regulators, including DNA-binding transcription factors, coactivators, corepressors and proteins that alter epigenetic modifications of DNA and nucleosomal histones, combine to influence the function of minimal promoters[2–4]. These transcriptional regulators act through a variety of DNA regulatory elements including enhancers, silencers and insulators, which are often located far from the gene promoters[2,3,5,6]. Typically, each DNA regulatory element binds a different subset of transcriptional regulators and thus is independently controlled, imparting modularity to gene regulation. The modular architecture of proteins and DNA is likely to be responsible for the complexity, versatility, flexibility and robustness of organisms and for their continued ability to evolve and adapt[7–9].

Biologists have studied modular genomic regulatory elements for about 30 years, but the availability of fully sequenced genomes has simplified functional analyses and greatly expanded their scope. As mentioned above, DNA-based regulatory modules can be widely dispersed in the genome; they may be located in the introns of the regulated gene, a few hundred base pairs to hundreds of kilobases 5′ or 3′ of the gene, or even in the introns of a neighboring irrelevant gene[5,6,10]. Before whole-genome sequences became available, it was necessary to test increasingly large segments of DNA in functional experiments, typically for their ability to drive gene expression in transgenic mice, to define the minimal segment of chromosome that most closely approximated the regulation of a gene *in vivo*. Although this criterion is still valid, the availability of complete genomes and the consequent detailed knowledge of the chromosomal milieu in which a gene is located mean that the location and size of potential regulatory domains can be readily visualized and biological analyses can be done in a more focused way.

Typically, genome sequence comparisons are used to detect noncoding genomic regions that have been evolutionarily conserved, presumably to maintain some critical biological function. Such regions, which have been called conserved noncoding sequences (CNSs), often correspond to dispersed transcriptional regulatory elements[11,12]. Knowledge of the genomic location of CNS regions greatly facilitates analysis of their biological functions; the regions may be cloned and tested in cell-based reporter assays or deleted or mutated in their natural genomic context or in the context of bacterial or yeast artificial chromosome transgenes. CNS regions may regulate a broad array of biological functions, not necessarily confined to transcriptional regulation. For example, CNSs may correspond to loci for noncoding RNAs or provide signals necessary for regulated mRNA splicing. These alternate functions will not be addressed here.

Several comparative genomics programs are available as freely accessible servers online, so that even someone who has access only to a public terminal can investigate loci of interest. The programs have been reviewed[12–16], but many 'bench scientists' are unfamiliar with the programs and might welcome a detailed introduction. The following sections provide some of the background necessary to use the programs knowledgeably as well as a brief discussion of the biological approaches needed to determine whether a specific CNS functions to regulate gene expression. The accompanying tutorial (**Supplementary Tutorial** online) provides a step-by-step online guide through the process of retrieving genomic sequences, creating annotations and using comparative genomics programs to identify CNS regions.

IL4: a case study

For many genes of immunological interest, for example the loci encoding interleukin 4 (*IL4*)[11,17–24], interferon-γ (*IFNG*)[17,25], the interleukin 2 receptor α (*IL2RA*)[26], stem cell leukemia (*SCL*)[27], lymphoblastic leukemia-1

Julie Nardone, K. Mark Ansel and Anjana Rao are with the Department of Pathology, Harvard Medical School and the CBR Institute for Biomedical Research, Boston, Massachusetts 02115, USA. Dong U. Lee is with the Department of Pathology and the Graduate Program in Immunology, Harvard Medical School, and the CBR Institute for Biomedical Research, Boston, Massachusetts 02115, USA.
e-mail: arao@cbr.med.harvard.edu

(*LYL1*)[28] and T cell receptor α, β, γ and δ[29], CNS regions contain functional gene regulatory elements. We chose the *IL4* locus as a case study because most if not all CNSs in this locus correspond to functional regulatory regions[11,17–22,30].

IL4 is part of the T helper type 2 (T_H2) cytokine cluster that also contains the coregulated genes encoding the cytokines interleukin 13 (IL-13) and IL-5 (*IL13* and *IL5*; **Fig. 1**) These three cytokine genes have similar expression patterns in T_H2 cells and in certain other immune cell types (natural killer T cells[31], mast cells[32] and basophils[33]). They are located in a larger locus that also includes the genes encoding the DNA repair protein RAD50, the kinesin family member Kif3a and the interferon-induced transcription factor IRF-1 (ref. 34; **Fig. 1a**).

IL4 is on mouse chromosome 11; the *IL4* exons are shown here in relation to regions in the mouse genome that are conserved relative to human (**Fig. 1b–d**). As described below, slightly different methods of measuring sequence conservation are used here, but larger values along the vertical axes always indicate more conservation (**Fig. 1b–d**). The sequences of the *IL4* exons are unexpectedly poorly conserved. Although this is unusual for most genes, cytokine genes are more divergent in their coding sequences than are other genes, perhaps because each species has its particular set of pathogens to which the immune system must adapt. The conserved regions that do not coincide with exons are CNS regions.

Notably, most of the CNS regions (**Fig. 1b–d**) correspond well to functionally defined regions[11,20–22,30]. Physical indicators of biological activity, such as DNase I hypersensitive sites[17–19] and sites of differential DNA methylation[35], have been identified in the *IL4* locus[23,24] (hypersensitive sites are discussed further below). Hypersensitive site I (**Fig. 1e**) corresponds to the CNS region at the *IL4* promoter, which is strongly conserved in evolution; hypersensitive site II corresponds to an intronic CNS and enhancer element that is active in T cells and mast cells[20,22,36]; and hypersensitive sites 1 and 2 are contained in CNS1, an enhancer element whose genomic deletion impairs both *IL4* and *IL13* expression[11,21]. Although the correlation between CNS regions and hypersensitive sites is notable, it is not absolute; in particular, there is no obvious CNS corresponding to hypersensitive site V_A, an inducible hypersensitive site and 3′ *IL4* enhancer that functions both *in vitro*[19] and *in vivo*[22,30].

Bioinformatic analysis

Two main assumptions underlie the utility of comparative genomic analyses for discovery of regulatory regions: that some function of regulatory elements will be conserved between species, and that functional conservation will be reflected in similar nucleotide sequences. There is no assumption as to which specific functions have been conserved. Selection in CNS areas could act to stabilize or change sequences that temporally, spatially or quantitatively control transcription by determining transcription factor binding sites, methylation status, DNA structure, arrangement in nucleosomes or some combination of these factors. Not enough is known about any locus to decide on the basis of sequence similarity which of these functions has been selected. For this reason, bioinformatic prediction of regulatory regions cannot be more than an adjunct to biological experiments at present.

With these caveats, is it worthwhile to identify CNS regions? Empirically, CNS regions correlate with the DNA modules (that is, the promoters, enhancers and repressor elements) that create and modify transcriptional specificity[5,6,37–39]. Although specific binding sites within the modules might not be conserved, the CNS regions delineate a manageable 'space' in which to search for notable motifs and test their functional importance in biological assays. This space will be more stringently defined as more full genome sequences (such as dog and chicken genomes) become available, allowing multiple genomic alignments of phylogenetically diverse species.

Identifying orthologs

The first step in finding CNS regions near an 'interesting' gene is to identify its ortholog in another species. That is, a pair of genes must be identified that have been 'derived from the same gene in a common ancestor'[40], in which one member of the pair is the gene of interest

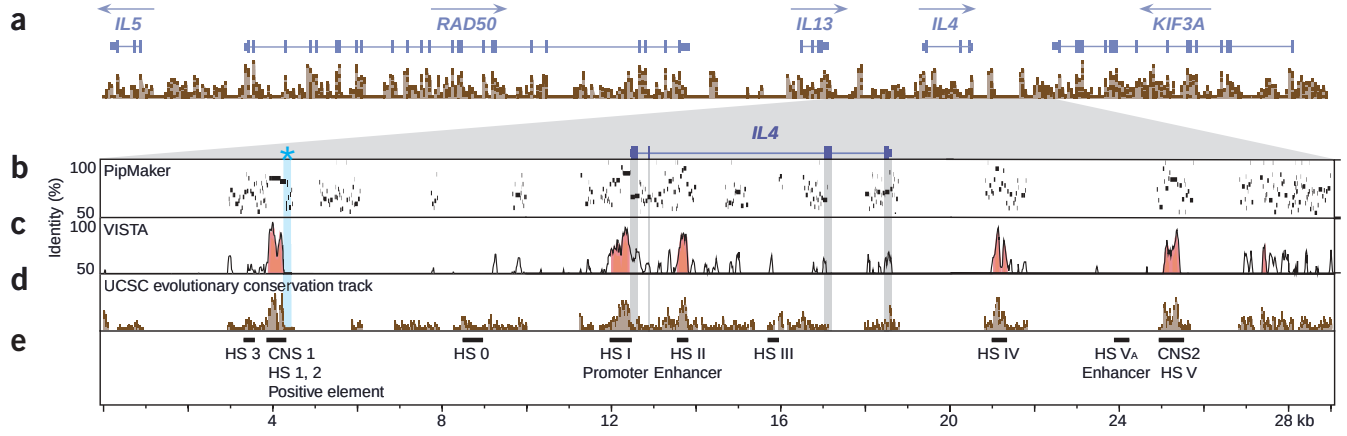

Figure 1 T_H2 cytokine locus on mouse chromosome 11. (**a**) Genes encoding three cytokines (*IL5*, *IL13* and *IL4*) are interspersed with those encoding Rad50 and Kif3a. Adapted from the University of California at Santa Cruz (UCSC) Genome Browser[45]. Nucleotide sequence, horizontal axis. *Rad50*, *IL13* and *IL4* are transcribed from left to right here, and *IL5* and *Kif3a* are transcribed from right to left. The height of the brown peaks is a measure of the probability that the underlying sequence is conserved, rather than simply residing in a slowly evolving section of the chromosome. (**b–d**) Measures of conservation in the *IL4* locus. Above, *IL4* and its exons (blue boxes and gray bars; includes 28 kb of the surrounding region). The bar and asterisk at approximately 4 kilobases (kb) indicate the region selected for the alignments of **Figure 3**. (**b**,**c**) Mouse and human *IL4* sequences, aligned with BLASTZ[55] and displayed with PipMaker[56] (**b**) or aligned with AVID[52] and displayed with VISTA[53] (**c**). Vertical axes, percent identity between the sequences. (**d**) Mouse and human conservation in the evolutionary conservation track of the UCSC Genome Browser[45,77], as described in **a**. (**e**) Black horizontal bars indicate the positions of functional regulatory regions. HS, hypersensitivity site.

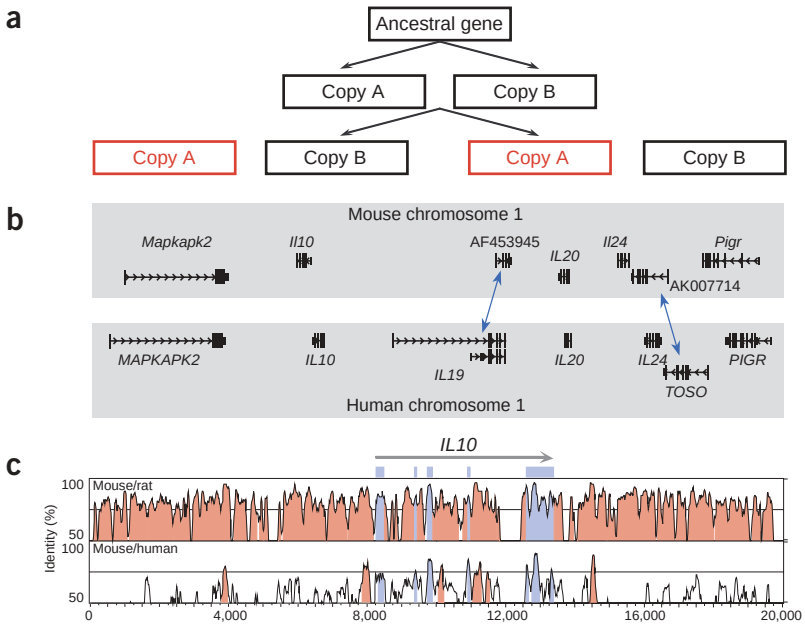

Figure 2 Identification of orthologs. (**a**) Orthologs and paralogs (modified from ref. 40 with permission from Kluwer Academic Publishers). Copy A in species 1 and copy A in species 2 are orthologs (orange). Homologous genes within a species (for example, copies A and B in species 1) are paralogs. (**b**) Orthologous genes in the class II cytokine clusters of human and mouse (modified from the UCSC Genome Browser[45]). Orthologs can be distinguished by their order in the genome. In this case, each cytokine is also the reciprocal best alignment of its ortholog. For the two loci in the mouse genome for which annotation is unavailable (AF453945 and AK007714), this reinforces the evidence for orthology given by position. (**c**) Comparison of mouse *IL10* genomic DNA with rat and human *IL10*. Vertical axis, percent identity with each organism as a function of the mouse sequence. Blue peaks are *IL10* exons; orange peaks are CNSs. Top: Mouse and rat, closely related organisms, have strong sequence similarity. This comparison would be used to identify small areas that are not conserved and are either not important or have been selected for different functions in the two species. The area of apparent nonconservation at approximately 12,000 base pairs (bp) is the result of missing rat sequence; the degree of conservation cannot be determined there. Bottom: In contrast, a comparison between mouse and human shows that there is much less conservation at this genetic distance. There is 'strong' CNS at the *IL10* promoter. As only a few other CNSs emerge, they can be assessed for biological function.

in the species chosen for functional studies. If gene duplication has produced a family of genes with similar sequences, the sequences being compared should be equivalent members. In **Figure 2a**, copy A in species 1 and copy A in species 2 are orthologs, whereas copies A and B in species 1 are homologous genes within the species and are paralogs. Although some regulatory regions may be conserved between copy A of species 1 and copy B of species 2 in the figure, more sequence and more function conservation between the orthologs would be expected.

A good indicator of orthology between genes is that flanking genes are conserved, in the same order, in the species being compared[41]. The class II cytokine cluster serves as an example (**Fig. 2b**). *IL10*, *IL19*, *IL20* and *IL24* are likely to have been the product of tandem gene duplications[42], but they are distinguishable in both mouse and human by their position relative to the flanking genes *MAPKAPK2* and *PIGR*. Genome browsers provide a clear view of a gene's chromosomal context and are useful both in determining orthology and for planning the extent of DNA to cover with functional assays. ENSEMBL[43] (European Molecular Biology Laboratory–European Bioinformatics Institute and the Sanger Institute), Map Viewer[44] (National Center for Biotechnology Information) and the University of California at Santa Cruz Genome Browser[45] are the main browsers now in use. In addition to showing flanking genes, the browsers can also show the locations of other useful biological entities, such as alternatively spliced exons, repeat sequences and bacterial artificial chromosome clones.

A frequently used measure of orthology, particularly between genes from evolutionarily distant species, is reciprocal best alignment. Again using the class II cytokine locus as an example (**Fig. 2b**), it is apparent that one of the cytokine genes is not accompanied by annotation in the mouse database. Use of human IL-19 in a BLAST search of the mouse proteome shows the protein encoded by AF453945 has the highest-scoring alignment; conversely, human IL-19 is the highest-scoring alignment in a BLAST search of the human proteome with the protein encoded by AF453945. Thus, these proteins are reciprocal best alignments and are therefore likely to be orthologs. Selection usually acts on a protein rather than on its coding sequence; as a result, protein searches are more sensitive than DNA searches, which is why they are used here. Although the method of reciprocal best alignments is useful in cases in which two organisms have diverged to the extent that orthologous genes now reside among different neighbors, it is not always accurate[46]. If a gene is the product of a recent duplication or has changed in function between the species, the reciprocal best alignment method may incorrectly identify paralogs as orthologs[47] (example, **Supplementary Tutorial** online).

Choice of species

What are the factors determining which species should be compared? This is discussed in detail in the accompanying tutorial (**Supplementary Tutorial** online), but the main points are as follows. Comparative genomics is robust in identifying regulatory regions and can be used successfully by pairing one species with any of several other species with different divergences[48]. However, because of the intrinsic limitations of the alignment programs (discussed below), regulatory regions that have been rearranged in the genome may be missed. For example, as mentioned earlier, the hypersensitive site VA enhancer in the *IL4* locus is not detected as a CNS by mouse-human sequence comparisons (**Fig. 1**). Comparisons of closely related species, such as mouse and rat or human and chimpanzee, will identify the regions where divergence is most readily tolerated by highlighting differences rather than similarities between sequences (**Fig. 2c**, top), whereas comparisons of distantly related organisms, such as mouse and chicken, will identify highly constrained sequences[49]. Sequence comparison of moderately related species, such as human and mouse, is ideal for a survey analysis such as this (**Fig. 2c**, bottom). Using multiple species for comparison greatly increases the power of the technique[48], allowing the boundaries of the CNS region to be further refined.

Making alignments

Once an ortholog has been found, the two sequences must be aligned. Sequence comparison programs use local or global alignment strategies or a combination of the two[50]. Global alignment programs produce a single alignment that is optimized across the entire length of two sequences. Local alignment programs find all of the high-quality alignments between two sequences, regardless of order or orientation. The goal of both strategies is to identify segments of the sequences being compared that are derived from the same ancestral sequence.

Because ancestral sequences are not available, it is difficult to evaluate the accuracy of the alignment strategies. One comparison of local and global alignment programs used test sequences that had been derived by computationally simulated evolution from the same 'ancestral sequence'[51]. In those conditions, global methods were able to detect more correct alignments than were local methods; however, nearly all of the alignments detected with local methods were accurate. It was noted, however, that both genome organization and the pattern of sequence evolution differ between drosophila, the model for evolution used in the study, and mammals[51]. Thus, until simulated sequences from other animals have also been tested, both global and local alignment programs should be used and the results should be compared.

Several alignment programs have publicly available online interfaces. AVID[52] (available through the VISTA[53] site online) and LAGAN[54] are global alignment programs. BLASTZ[55] (available through the PipMaker[56] site online) is a local alignment program. Dialign-Chaos[57] uses combined global and local alignment strategies. All of these programs accept two or more sequences for comparison and return alignment results by e-mail. Detailed instructions for submitting sequences to the VISTA and PipMaker websites are available (**Supplementary Tutorial** online).

Practical considerations

As expected, the local and global alignment strategies give different results[39]. Positioning of gaps and the nucleotides adjacent to gaps are particularly sensitive to method (BLASTZ and AVID alignments, **Fig. 3a**,**b**). A practical consequence of these differences is apparent by comparison of the boxed residues in the alignments; these are potential binding sites for the transcription factor NFAT. In the AVID alignment (**Fig. 3b**), the site does not seem to be conserved. In the BLASTZ alignment (**Fig. 3a**), the site is detectably but not perfectly conserved.

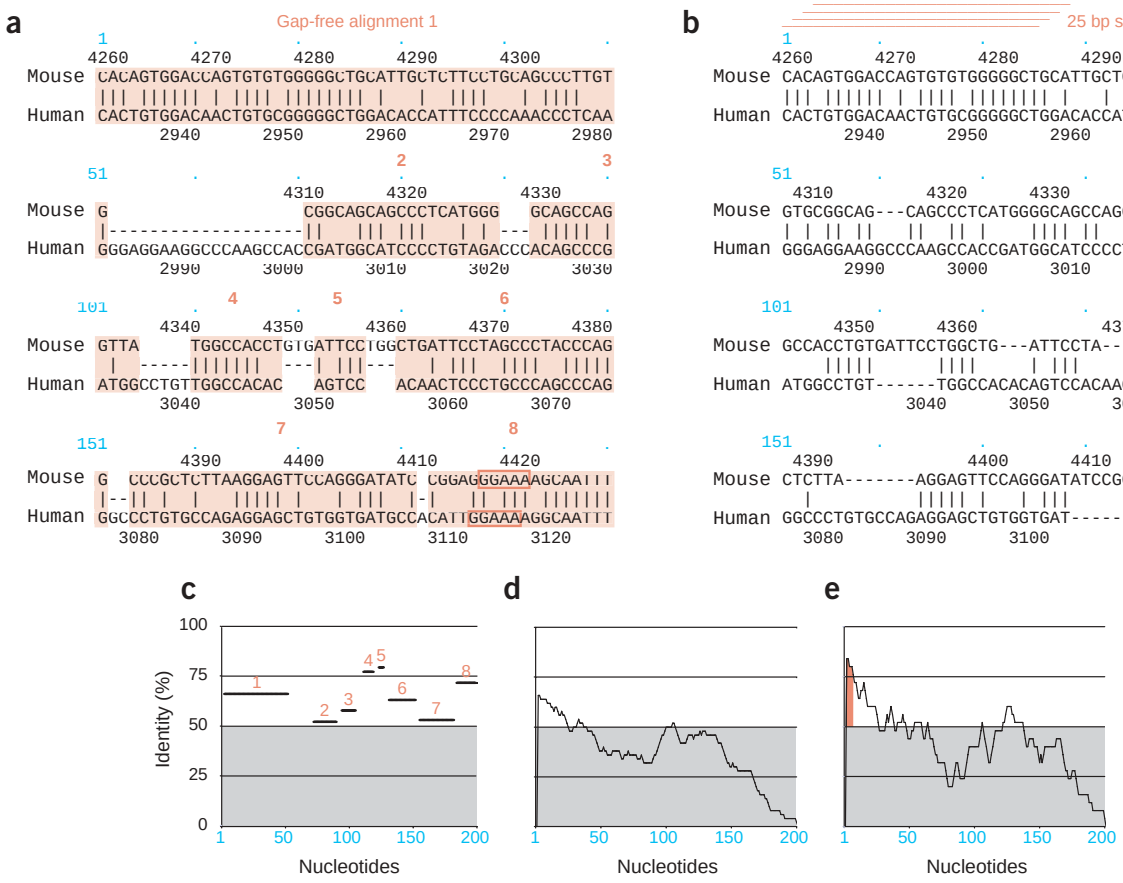

Figure 3 Alignment of a section of the *IL4* locus on mouse chromosome 11 with the corresponding region of human chromosome 5. Comparison of alignment and display methods (region analyzed, **Fig. 1**). (**a**) Excerpt of a BLASTZ-PipMaker alignment of the human and mouse *IL4* loci. Top, mouse sequence; bottom, human. The program computes percent sequence identity within gap-free aligned regions; the aligned segments visible in this excerpt are shaded in orange and numbered in orange. Numbering for the mouse sequence corresponds to that used as the horizontal axes in **c**–**e**; numbering for the alignment is in blue. (**b**) Excerpt of an AVID-VISTA alignment of the same sequences as in **a**. The program computes percent sequence identity within sliding windows; four sample 25–base pair windows are in orange above the alignment. The AVID-VISTA alignment splits a potential conserved NFAT site (boxed) near the end of the sequence, which is preserved (albeit offset) in the BLASTZ-PipMaker alignment in **a**. (**c**) PipMaker plot of BLASTZ alignment in **a**. The numbering of the alignment in **a** serves as the horizontal axis. (**d**) VISTA plot of AVID alignment with a 50–base pair sliding window. The numbering of the alignment in **b** serves as the horizontal axis. (**e**) VISTA plot of AVID alignment with a 25–base pair sliding window.

In addition to providing online portals to the alignment programs, VISTA and PipMaker generate graphical representations of the alignments. Both programs plot the percent identity between the aligned sequences as a function of one of the sequences, chosen as the 'base' or horizontal axis. The two programs use different methods to calculate the percent identity. PipMaker shows the positions of aligned segments that may contain mismatches but do not contain gaps (**Fig. 3c**, PipMaker-style plot of the sequence comparison in **Fig. 3a**). The data are presented as a series of bars, with the length of each bar representing the length of the segment and the vertical axis indicating the percent identity between the two sequences within the segment. The range of PipMaker's vertical axis is 50–100% identity. In contrast, VISTA shows the percent identity between two sequences by sliding a fixed 'window' over an alignment (**Fig. 3b**). The window size is set by the user, as is the offset between one window and the next (**Fig. 3d**,**e**, window sizes of 50 and 25 base pairs, respectively, and an offset of 1 base pair). Percent identity is plotted as a function of position within the alignment. With the default parameters of VISTA, the range of the vertical axis is 50–100% identity. Peaks higher than 75% are presented in color and are considered to be CNS regions. A smaller window size results in more and higher peaks but may also increase background.

Functional assessment

Once CNS regions have been identified, the next step is to evaluate their function(s) in biological experiments. Of the many techniques available to assess regulatory function[3], only those that are likely to be the most informative are discussed here. Given the lacunae in knowledge about these at present, no combination of computational analyses can substitute for functional assessment. Without 'wet-lab' experiments, the biological functions of modules that have already been identified cannot be predicted, nor can it be confirmed that a specific module actually affects a given gene.

DNase I hypersensitivity

A useful first approach is to ask whether the CNS regions identified by bioinformatic analysis correspond to DNase I–hypersensitive sites in the cell type of interest, either in resting conditions or after appropriate stimulation. Hypersensitive sites are regions at which local nucleosome organization has been altered from that of surrounding areas; they are often occupied by or show increased accessibility to transcription factors and other DNA-binding proteins[58]. Hypersensitive site mapping is ideal for rapid survey of CNS sequences in large genomic regions, as the initial bioinformatic analysis immediately pinpoints the optimal restriction digests needed for the assay. Ideally, bioinformatic and hypersensitive site analyses should extend for 50–100 kilobases in either direction from the gene, and intronic regions of neighboring genes should be included in the analysis, as they may contain important regulatory elements[10,59,60]. As mentioned above, not all hypersensitive sites correspond to CNS regions and, conversely, any given CNS region may be hypersensitive to DNase I only in certain cell types or developmental stages or in particular conditions of stimulation.

Once hypersensitive site mapping has been completed for a cell type of interest, it is worthwhile to extend it to other cell types and stimulation conditions. The observed patterns may lead to specific predictions regarding function, which can be tested by judicious combinations of functional studies, as described below. The promoter is usually a CNS sequence and will obviously be biologically relevant; its identification depends less on bioinformatic analysis than on accurate mapping of the transcription start site. For genes with higher transcription after stimulation, CNS regions that show increased hypersensitivity in stimulated cells may correspond to inducible enhancers whose function can be tested in standard reporter assays[19,25,26].

Developmental and cell lineage analyses are particularly informative. For example, *IL4* is expressed by T_H2 cells of lymphoid lineage as well as by mast cells, which are of the myeloid lineage. It is silenced not only in cells of unrelated lineages, such as fibroblasts, but also in T_H1 cells, which derive from the same naive precursor T cells as do T_H2 cells. In addition, *IL4* is expressed by a small subset of myotubes, where it facilitates myotube-myoblast fusion during muscle growth[61]. The different *IL4*-expressing and nonexpressing cell types are likely to display different patterns of hypersensitive sites, whose functions in each cellular context may be surmised based on these expression patterns. T_H1 and T_H2 cells show differences in hypersensitive site patterns and histone acetylation throughout the region of the *IL4* locus shown in **Figure 1**, but these differences end just before the start of the neighboring (3′) *KIF3A* gene[62]. A potentially testable prediction, therefore, is that CNS2 (hypersensitive site V) in *IL4* contains an insulator element[63] that ensures that *KIF3A* is not differentially expressed. In another example, precursor naive T cells show DNase I hypersensitivity only at hypersensitive sites 3 and IV, located 5′ and 3′ of *IL4*, respectively[24]. T_H1 cells, which derive from the same naive precursors but have silenced *IL4*, continue to display hypersensitive site 3 (ref. 18) and hypersensitive site IV (refs. 17,19). However, neither of these sites is apparent in fibroblasts, an unrelated cell type that has also silenced *IL4*. A testable hypothesis following from these data is that hypersensitive sites 3 and IV have distinct, cell type–specific functions: in T_H1 cells, the sites may participate in silencing of the cytokine genes, whereas in naive T cells, the sites may be responsible for the 'poised' state of the locus, for which the exact stimulation conditions determine whether gene activation or silencing prevails.

In the near future, high-throughput methods of surveying hypersensitive sites will be available, as demonstrated by a pilot study[64]. At this early stage, specificity, scale and cost have yet to be optimized. For example, although that study[64] used human $CD4^+$ T cells, it did not show the hypersensitive sites 3 and IV that were defined in the mouse. Nonetheless, it is likely that global surveys of entire genomes will become a valuable tool for the discovery of regulatory regions, especially if the analysis is done systematically in different species with a variety of different cell types, developmental stages and stimulation conditions.

Targeted disruption

Undoubtedly the most reliable means of assessing *in vivo* function is targeted disruption of putative regulatory regions. Deletions and mutations of regulatory regions can be done either in the native chromosomal context or in large bacterial or yeast artificial chromosome transgenes[11,21,22,60,65–67]. Deletion of positive regulatory elements such as enhancers would result in decreased gene expression; deletion of negative regulatory elements such as silencers would lead to increased gene expression in cells that either normally express or normally silence the gene; and loss of an insulator element may lead to inappropriate expression of either the gene in question or a neighboring gene in an irrelevant cell type. For immunologically relevant loci such as *IL4*, for which the ability to survive attack by pathogens is crucial for reproductive fitness and survival, deletion of individual hypersensitive site sites

may have only a partial effect[11,21,22], most likely because evolutionary pressures have imposed functional redundancy such that more than one regulatory region participates in gene activation or silencing. In such cases, multiple mutations may be needed to produce strong effects[65,66]. The function of the disrupted locus should be assessed in a variety of cell types, developmental stages and stimulation conditions, as a CNS region that seems to be nonfunctional in one condition may demonstrate a notable function in another.

Cell-based and biochemical assays

Although disruption of regulatory regions provides essential clues to *in vivo* function, the mechanism by which a given regulatory region influences transcription through its target promoter may be investigated with cell-based assays. A variety of reporter assays, using cell lines or transgenic animals, have been used to assess whether putative enhancer, silencer and insulator elements influence gene expression from target promoters[3]. Furthermore, as DNA-binding proteins associated with a regulatory region are likely to recruit DNA- and histone-modifying enzymes, a CNS involved in regulating gene expression is likely to be a focus for differential histone modifications or differential DNA methylation in cells that express or do not express the gene[25,35,62,68,69]. This can be evaluated by chromatin immunoprecipitation with antibodies to modified histone tails and by restriction digestion with methylation-sensitive enzymes.

The next step is to identify the relevant protein-binding sites in the CNS regions, as well as the proteins which actually bind to the sites. Bioinformatic analysis[12,70–72] is less useful here, in part because of the variation in results from different alignment programs (**Fig. 3**) and in part because the complex evolution of transcription factor binding diminishes the prospects for predicting individual sites. Transcription factors typically bind a span of five to eight base pairs of DNA, with substitution possible in at least one position[73]. Flexibility in recognition allows discrimination of sites by affinity and can impose a requirement for a partner binding protein if the affinity becomes sufficiently low. Often a promoter or other regulatory region contains multiple copies of binding sites, sometimes with different affinities and in proximity to binding sites for different factors. This redundancy, combined with the ease with which the short sites can be created in DNA sequence by chance, is a form of genetic buffering[74] and means that loss of a particular transcription factor binding site, possibly to a new function, does not necessarily lead to loss of regulation by that factor. In fact, the overall gene regulatory function of a conserved enhancer can be maintained in the face of substantial evolutionary changes in the order and affinity of transcription factor binding sites within the conserved module[37]. Reliable bioinformatic methods for predicting transcription factor binding sites in a CNS, while taking into consideration binding site turnover secondary to evolutionary drift, are still being developed.

Because of these difficulties, identification of transcription factors binding to a CNS region still relies on a 'candidate' approach. Simple methods, such as gel mobility-shift assays, have occasionally been successful[75], but the readout tends to be overwhelmed by ubiquitous factors abundant in nuclear extracts. One-hybrid assays have also yielded plausible candidates in a handful of cases[67,76]. Once transcription factor candidates have been identified, however, their *in vivo* functions can be tested with standard approaches, using chromatin immunoprecipitation to monitor *in vivo* binding of the transcription factor to the CNS regions in the expected cell type in the appropriate conditions of stimulation[19,25] and monitoring the effect of targeted deletion of the gene encoding the candidate transcription factor[22,66,67].

Tutorial

The accompanying tutorial (**Supplementary Tutorial** online) provides a road map to the identification of CNS regions. It includes a discussion of the choice of species to be compared, expands the explanation of orthologs and lists the programs used as well as additional online resources available. The first section covers how to define the regions to analyze and retrieve the sequences; the next, how to determine the transcripts in this region and use them to annotate the genomic sequences; and the last, how to compare the genomic sequences of two species with both VISTA and PipMaker and interpret the results.

Conclusion

Bioinformatic approaches can be extremely valuable in identifying evolutionarily conserved CNS regions that correspond to cell type–specific, inducible or developmentally important regulatory elements. Comparative genomics does not predict specific function, but it also does not depend on the existence of a particular function, so it is not biased against regions that are used for previously unknown regulatory purposes. A 'strong' CNS sequence with an unknown function can be used in cell-based assays or transgenic animals to discover new control processes.

Note: Supplementary information is available on the Nature Immunology website.

1. Baltimore, D. Our genome unveiled. *Nature* **409**, 814–816 (2001).
2. Levine, M. & Tjian, R. Transcription regulation and animal diversity. *Nature* **424**, 147–151 (2003).
3. Carey, M. & Smale, S.T. *Transcriptional Regulation in Eukaryotes: Concepts, Strategies, and Techniques* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 2000).
4. Fischle, W., Wang, Y. & Allis, C.D. Histone and chromatin cross-talk. *Current Opinion in Cell Biology* **15**, 172–183 (2003).
5. Arnone, M. & Davidson, E. The hardwiring of development: organization and function of genomic regulatory systems. *Development* **124**, 1851–1864 (1997).
6. Davidson, E.H. Genomic Regulatory Systems: Development and Evolution (Academic, San Diego, 2001).
7. Kirschner, M. & Gerhart, J. Evolvability. *Proc. Natl. Acad. Sci. USA* **95**, 8420–8427 (1998).
8. Locascio, A., Manzanares, M., Blanco, M.J. & Nieto, M.A. Modularity and reshuffling of Snail and Slug expression during vertebrate evolution. *Proc. Natl. Acad. Sci. USA* **99**, 16841–16846 (2002).
9. Lynch, M. & Conery, J.S. The origins of genome complexity. *Science* **302**, 1401–1404 (2003).
10. Mancini-DiNardo, D., Steele, S.J.S., Ingram, R.S. & Tilghman, S.M. A differentially methylated region within the gene *Kcnq1* functions as an imprinted promoter and silencer. *Hum. Mol. Genet.* **12**, 283–294 (2003).
11. 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).
12. Loots, G.G., Ovcharenko, I., Pachter, L., Dubchak, I. & Rubin, E.M. rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res.* **12**, 832–839 (2002).
13. Pennacchio, L.A. & Rubin, E.M. Genomic strategies to identify mammalian regulatory sequences. *Nat. Rev. Genet.* **2**, 100–109 (2001).
14. Frazer, K.A., Elnitski, L., Church, D.M., Dubchak, I. & Hardison, R.C. Cross-species sequence comparisons: a review of methods and available resources. *Genome Res.* **13**, 1–12 (2003).
15. Pennacchio, L.A. & Rubin, E.M. Comparative genomic tools and databases: providing insights into the human genome. *J. Clin. Invest.* **111**, 1099–1106 (2003).
16. Wasserman, W.W. & Sandelin, A. Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.* **5**, 276–287 (2004).
17. Agarwal, S. & Rao, A. Modulation of chromatin structure regulates cytokine gene expression during T cell differentiation. *Immunity* **9**, 765–775 (1998).
18. Takemoto, N. *et al.* Th2-specific DNase I-hypersensitive sites in the murine IL-13 and IL-4 intergenic region. *Int. Immunol.* **10**, 1981–1985 (1998).
19. Agarwal, S., Avni, O. & Rao, A. Cell-type-restricted binding of the transcription factor NFAT to a distal IL-4 enhancer in vivo. *Immunity* **12**, 643–652 (2000).
20. Lee, G.R., Fields, P.E. & Flavell, R.A. Regulation of IL-4 gene expression by distal regulatory elements and GATA-3 at the chromatin level. *Immunity* **14**, 447–459 (2001).
21. Mohrs, M. *et al.* Deletion of a coordinate regulator of type 2 cytokine expression in mice. *Nat. Immunol.* **2**, 842–847 (2001).
22. Solymar, D.C., Agarwal, S., Bassing, C.H., Alt, F.W. & Rao, A. A 3′ enhancer in the *IL-4* gene regulates cytokine production by Th2 cells and mast cells. *Immunity* **17**, 41–50 (2002).
23. Smale, S.T. & Fisher, A.G. Chromatin structure and gene regulation in the immune system. *Annu. Rev. Immunol.* **20**, 427–462 (2002).

24. Ansel, K.M., Lee, D.U. & Rao, A. An epigenetic view of helper T cell differentiation. *Nat. Immunol.* **4**, 616–623 (2003).
25. Lee, D.U., Avni, O., Chen, L. & Rao, A. A distal enhancer in the interferon-γ (*IFN-γ*) locus revealed by genome sequence comparison. *J. Biol. Chem.* **279**, 4802–4810 (2004).
26. Kim, H.P., Kelly, J. & Leonard, W.J. The basis for IL-2-induced IL-2 receptor α chain gene regulation: importance of two widely separated IL-2 response elements. *Immunity* **15**, 159–172 (2001).
27. Göttgens, B. *et al.* Long-range comparison of human and mouse *SCL* loci: localized regions of sensitivity to restriction endonucleases correspond precisely with peaks of conserved noncoding sequences. *Genome Res.* **11**, 87–97 (2001).
28. Chapman, M.A. *et al.* Comparative and functional analyses of *LYL1* loci establish marsupial sequences as a model for phylogenetic footprinting. *Genomics* **81**, 249–259 (2003).
29. Glusman, G. *et al.* Comparative genomics of the human and mouse T cell receptor loci. *Immunity* **15**, 337–349 (2001).
30. Amsen, D. *et al.* Instruction of distinct CD4 T helper cell fates by different notch ligands on antigen-presenting cells. *Cell* **117**, 515–526 (2004).
31. Hammond, K.J. & Kronenberg, M. Natural killer T cells: natural or unnatural regulators of autoimmunity? *Curr. Opin. Immunol.* **15**, 683–689 (2003).
32. Weiss, D.L. & Brown, M.A. Regulation of IL-4 production in mast cells: a paradigm for cell-type-specific gene expression. *Immunol. Rev.* **179**, 35–47 (2001).
33. Falcone, F.H., Haas, H. & Gibbs, B.F. The human basophil: a new appreciation of its role in immune responses. *Blood* **96**, 4028–4038 (2000).
34. Frazer, K.A. *et al.* Computational and biological analysis of 680 kb of DNA sequence from the human 5q31 cytokine gene cluster region. *Genome Res.* **7**, 495–512 (1997).
35. Lee, D.U., Agarwal, S. & Rao, A. Th2 lineage commitment and efficient IL-4 production involves extended demethylation of the *IL-4* gene. *Immunity* **16**, 649–660 (2002).
36. Hural, J.A., Kwan, M., Henkel, G., Hock, M.B. & Brown, M.A. An intron transcriptional enhancer element regulates IL-4 gene locus accessibility in mast cells. *J. Immunol.* **165**, 3239–3249 (2000).
37. Ludwig, M.Z., Bergman, C., Patel, N.H. & Kreitman, M. Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* **403**, 564–567 (2000).
38. Stern, D.L. Evolutionary developmental biology and the problem of variation. *Evolution* **54**, 1079–1091 (2000).
39. Bergman, C.M. & Kreitman, M. Analysis of conserved noncoding DNA in drosophila reveals similar constraints in intergenic and intronic sequences. *Genome Res.* **11**, 1335–1345 (2001).
40. Doyle, J.J. & Gaut, B.S. Evolution of genes and taxa: a primer. *Plant Mol. Biology* **42**, 1–23 (2000).
41. Wolfe, K.H. & Shields, D.C. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* **387**, 708–713 (1997).
42. Lutfalla, G. *et al.* Comparative genomic analysis reveals independent expansion of a lineage-specific gene family in vertebrates: The class II cytokine receptors and their ligands in mammals and fish. *BMC Genomics* **4**, 29 (2003).
43. Birney, E. *et al.* An overview of Ensembl. *Genome Res.* **14**, 925–928 (2004).
44. Wheeler, D.L. *et al.* Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res.* **32**, D35–40 (2004).
45. Karolchik, D. *et al.* The UCSC genome browser database. *Nucleic Acids Res.* **31**, 51–54 (2003).
46. Koski, L.B. & Golding, G.B. The closest BLAST hit is often not the nearest neighbor. *J. Mol. Evol.* **52**, 540–542 (2001).
47. Forsyth, S., Horvath, A. & Coughlin, P. A review and comparison of the murine α1-antitrypsin and α1-antichymotrypsin multigene clusters with the human clade A serpins. *Genomics* **81**, 336–345 (2003).
48. Thomas, J.W. *et al.* Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* **424**, 788–793 (2003).
49. Cooper, G.M. *et al.* Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. *Genome Res.* **13**, 813–820 (2003).
50. Durbin, R., Eddy, S.R., Krogh, A. & Mitchison, G. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids* (Cambridge University Press, Cambridge, 1998).
51. Pollard, D.A., Bergman, C.M., Stoye, J., Celniker, S.E. & Eisen, M.B. Benchmarking tools for the alignment of functional noncoding DNA. *BMC Bioinformatics* **5**, 6 (2004).
52. Bray, N., Dubchak, I. & Pachter, L. AVID: A global alignment program. *Genome Res.* **13**, 97–102 (2003).
53. Mayor, C. *et al. VISTA:* visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics* **16**, 1046–1047 (2000).
54. Brudno, M. *et al.* LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Res.* **13**, 721–731 (2003).
55. Schwartz, S. *et al.* Human-mouse alignments with BLASTZ. *Genome Res.* **13**, 103–107 (2003).
56. Schwartz, S. *et al.* PipMaker—A web server for aligning two genomic DNA sequences. *Genome Res.* **10**, 577–586 (2000).
57. Brudno, M., Chapman, M., Gottgens, B., Batzoglou, S. & Morgenstern, B. Fast and sensitive multiple alignment of large genomic sequences. *BMC Bioinformatics* **4**, 66 (2003).
58. Gross, D.S. & Garrard, W.T. Nuclease hypersensitive sites in chromatin. *Annu. Rev. Biochem.* **57**, 159–197 (1988).
59. Adlam, M. & Siu, G. Hierarchical interactions control *CD4* gene expression during thymocyte development. *Immunity* **18**, 173–184 (2003).
60. Lee, G.R., Fields, P.E., Griffin, T.J. & Flavell, R.A. Regulation of the Th2 cytokine locus by a locus control region. *Immunity* **19**, 145–153 (2003).
61. Horsley, V., Jansen, K.M., Mills, S.T. & Pavlath, G.K. IL-4 acts as a myoblast recruitment factor during mammalian muscle growth. *Cell* **113**, 483–494 (2003).
62. Yamashita, M. *et al.* Identification of a conserved GATA3 response element upstream proximal from the interleukin-13 gene locus. *J. Biol. Chem.* **277**, 42399–42408 (2002).
63. Burgess-Beusse, B. *et al.* The insulation of genes from external enhancers and silencing chromatin. *Proc. Natl. Acad. Sci. USA* **99**, 16433–16437 (2002).
64. Crawford, G.E. *et al.* Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. *Proc. Natl. Acad. Sci. USA* **101**, 992–997 (2004).
65. Ellmeier, W., Sunshine, M.J., Maschek, R. & Littman, D.R. Combined deletion of CD8 locus *cis*-regulatory elements affects initiation but not maintenance of *CD8* expression. *Immunity* **16**, 623–634 (2002).
66. Taniuchi, I., Sunshine, M.J., Festenstein, R. & Littman, D.R. Evidence for distinct *CD4* silencer functions at different stages of thymocyte differentiation. *Mol. Cell* **10**, 1083–1096 (2002).
67. Taniuchi, I. *et al.* Differential requirements for Runx proteins in CD4 repression and epigenetic silencing during T lymphocyte development. *Cell* **111**, 621–633 (2002).
68. Avni, O. *et al.* T_H cell differentiation is accompanied by dynamic changes in histone acetylation of cytokine genes. *Nat. Immunol.* **3**, 643–651 (2002).
69. Bird, A. DNA methylation patterns and epigenetic memory. *Genes Dev.* **16**, 6–21 (2002).
70. Schug, J. & Overton, G.C. http://www.cbil.upenn.edu/tess (Computational Biology and Informatics Laboratory, School of Medicine, University of Pennsylvania, Philadelphia, 1997).
71. Kel-Margoulis, O.V. *et al.* Composition-sensitive analysis of the human genome for regulatory signals. *In Silico Biol.* **3**, 145–171 (2003).
72. Lenhard, B. *et al.* Identification of conserved regulatory elements by comparative genome analysis. *J. Biol.* **2**, 13.1–13.11 (2003).
73. Wray, G.A. *et al.* The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.* **20**, 1377–1419 (2003).
74. Rutherford, S.L. From genotype to phenotype: buffering mechanisms and the storage of genetic information. *Bioessays* **22**, 1095–1105 (2000).
75. Bell, A.C., West, A.G. & Felsenfeld, G. The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell* **98**, 387–396 (1999).
76. Szabo, S.J. *et al.* A novel transcription factor, T-bet, directs Th1 lineage commitment. *Cell* **100**, 655–669 (2000).
77. Hardison, R.C. Comparative genomics. *PLoS Biol.* **1**, E58 (2003).