Annotation of cis-regulatory elements by identification, subclassification, and functional assessment of multispecies conserved sequences

Jim R. Hughes*, Jan-Fang Cheng†, Nicki Ventress*, Shyam Prabhakar†, Kevin Clark*, Eduardo Anguita*, Marco De Gobbi*, Pieter de Jong‡, Eddy Rubin‡, and Douglas R. Higgs*§

*Medical Research Council Molecular Haematology Unit, Weatherall Institute of Molecular Medicine, John Radcliffe Hospital, Headington, Oxford OX3 9D5, United Kingdom; †Department of Energy Joint Genome Institute, Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720; and §Children’s Hospital and Research Center at Oakland, Oakland, CA 94609

Communicated by David Weatherall, University of Oxford, Oxford, United Kingdom, April 28, 2005 (received for review November 5, 2004)

An important step toward improving the annotation of the human genome is to identify cis-acting regulatory elements from primary DNA sequence. One approach is to compare sequences from multiple, divergent species. This approach distinguishes multispecies conserved sequences (MCSs) in noncoding regions from more rapidly evolving neutral DNA. Here, we have analyzed a region of ~238kb containing the human α-globin cluster that was sequenced and/or annotated across the syntenic region in 22 species spanning 500 million years of evolution. Using a variety of bioinformatic approaches and correlating the results with many aspects of chromosome structure and function in this region, we were able to identify and evaluate the importance of 24 individual MCSs. This approach sensitively and accurately identified previously characterized regulatory elements but also discovered unidentified promoters, exons, splicing, and transcriptional regulatory elements. Together, these studies demonstrate an integrated approach by which to identify, subclassify, and predict the potential importance of MCSs.

With the development of multiple computational algorithms and appropriate biological data, it is now possible to detect coding sequences relatively efficiently and accurately from primary DNA sequence (1). By contrast, it is still difficult to identify and assign function to noncoding sequences that include critical cis-acting regulatory elements such as promoters, enhancers, silencers, locus control elements, nuclear matrix attachment sites, and origins of replication. A promising approach is to search for conserved orthologous noncoding sequences from multiple, evolutionarily diverse species (2–8), so-called multispecies conserved sequences (MCSs). However, at present, it is not established how best to design and carry out such searches. Furthermore, even when complete, it is not clear how to evaluate and prioritize MCSs for functional analysis.

An important issue is to know which species to include in searches for MCSs. When comparing the sequences of distantly related species [e.g., birds and mammals, which diverged 310 million years (my) ago], cis elements may have diverged too much to be easily identified. By contrast, when comparing more closely related species [e.g., rodents and primates, which diverged 64–74 my ago], it may be difficult to distinguish between sequence homology resulting from a slow rate of evolution and the conservation of functionally important cis elements. Nevertheless, it is becoming clear that analyzing multiple species in different combinations adds significant power to the identification of functionally important cis-acting sequences (2, 3).

Even when MCS can be unequivocally identified, there is often no way to judge how sensitive the searches have been. Do these routines overcall or undercall regulatory elements? Furthermore, the true functional significance of MCSs often remains untested and/or unknown because there is usually very limited experimental data to link genome structure to function at the locus in question. Therefore, to establish accurate and sensitive routines for evaluating cis-regulatory sequences, it will be necessary to analyze well characterized regions of the human genome so that such sequences can be recognized and their potential importance can be assessed and prioritized for functional studies.

Here, we have sequenced and/or annotated and compared extensive regions (47–238 kb) of conserved syntenies containing the α-globin cluster, its known regulatory elements and flanking genes in 22 species representing widely divergent groups, spanning 500 my of evolution. Multiple alignments were analyzed with a variety of bioinformatic approaches, and the results were compared with algorithms for the detection of MCSs described in ref. 9 and a newly developed analysis tool, GUMBY (S. Prabhakar, unpublished data). The approach taken here allowed us to identify 24 MCSs within a 238-kb region containing and surrounding the human α-globin cluster. Previous studies of this locus have localized all DNAS1 hypersensitive sites (DHHSs; refs. 10 and 11), characterized its nuclear localization (12), nuclear matrix attachment (13), pattern of replication (14), chromatin structure (15), chromatin modification (15, 16) DNA methylation (17), and pattern of gene expression in normal loci and those with natural deletions (18), allowing us to evaluate in full the functional significance of the MCS elements identified here.

Materials and Methods

Clone Isolation and Sequencing. Bacterial artificial chromosomes (BACs) used in this study were isolated and separated by using standard techniques (19); all details concerning multiple alignments of the discussed elements are available as additional material from the authors on request. All BACs used in this study are available from the BACPAC Resources Centre (http://bacpac.org). The sequences of the chimpanzee, cow, and cat were obtained from the ENCODE project (http://genome.gov/ENCODE), region ENm008.

Sequence Assembly and Finishing. Sequence reads were base called by using PHRED 0.020425 and assembled by using PHRAP (www.phrap.org) into a GAP4 database. Standard finishing methods were used.

Abbreviations: uMRE, human α-globin major regulatory element; DHS, DNaseI hypersensitive site; MCS, multispecies conserved sequences; MCS-E, MCS-previously unidentified exon; MCS-P, MCS promoters; MCS-R, MCS-other transcriptional regulatory element; MCS-S, MCS-spooling element; my, million years; TFB, transcription factor binding site.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (species, clone names, and accession nos.): baboon, RP4I-9BL24, AC145461; colobus, CH32-179016, AC148220; ducky titi, LBS13F19, AC145465; marmoset, CH259-50K11, AC146591, and CH259-17702Q, AC145483; owl monkey, CH258-305M22, AC146782; squirrel monkey, CH254-2006G16, AC146463, and CH254-1671T1C, AC151838; lemurs, LBS2-9K66, AC145461; pig, RP46-21018, AC120214, and RP46-293223, AC120215; dog, RP81-84F12, AC120212, and RP81-397M1, AC120213; hedgehog, LBS4-35K82, AC150435; opossum, LB3-33C20, AC120504, and LB3-206D4, AC139599, and LB3-17112, AC148752.

To whom correspondence should be addressed. E-mail: doug.higgs@imm.ox.ac.uk.

© 2005 by The National Academy of Sciences of the USA
Sequence Annotation and Multiple Alignment. Sequences were annotated against EST data when available for that species (http://genom.ucsc.edu and www.ensembl.org) or by using cross-species comparison of available ESTs (MACVECTOR 7.2, Accelrys, Inc., San Diego). The presence of non-syntenic features were evaluated by using a modification of HPREP and stored in AceDB (www.acedb.org). Repeat annotation was generated by using REPEATMASKER (www.repeatmasker.org). Multiple alignments were generated and/or visualized by using VISTA (www-gsd.lbl.gov/vista/index.shtml); MULTIPIPMAKER (http://pipmaker.bx.psu.edu/pipmaker), and MULTILAGAN (http://lagan.stanford.edu/lagan_web/index.shtml). Multiple alignments were analyzed as described in Results and were also compared with data obtained by using WEBMCS (http://research.nhgrl.nih.gov/MCS/submit.shtml). In addition, MULTILAGAN alignments were analyzed for noncoding conservation by using the algorithm GUMBY, a tool for detecting statistically significant conserved regions in pairwise or multiple alignments of DNA sequences at any alignable evolutionary distance. Individual MCSs were aligned by using CLUSTALW (MACVECTOR 7.2; Accelrys); BESTFIT (WISCONSIN PACKAGE 10.3, GCG) and DIALIGN (www.genomatix.de/cgi-bin/dialign/dialign.pl) and manually optimized.

Sequence Analysis. Potential transcription factor binding sites (TFBs) were identified by using TRANSFAC 6.0 (www.gene-regulation.com); TRES (http://bioportal.bic.mus.edu/tres), and REGULATORY VISTA (http://gsd.lbl.gov/vista/rvista/submit.shtml). Plots of Observed/Expected for the CpG dinucleotide were generated by using WINDOW and STATPLOT 10.3 (GCC) comparing and were also compared with data obtained by using WEBMCS (http://research.nhgrl.nih.gov/MCS/submit.shtml). In addition, MULTILAGAN alignments were analyzed for noncoding conservation by using the algorithm GUMBY, a tool for detecting statistically significant conserved regions in pairwise or multiple alignments of DNA sequences at any alignable evolutionary distance. Individual MCSs were aligned by using CLUSTALW (MACVECTOR 7.2; Accelrys); BESTFIT (WISCONSIN PACKAGE 10.3, GCG) and DIALIGN (www.genomatix.de/cgi-bin/dialign/dialign.pl) and manually optimized.

Results

A Region of Conserved Synteny Maintained Throughout 500 my of Evolution. The structure of the α-globin cluster was established by DNA sequence analysis (6- to 8-fold coverage) in 22 divergent species representing well defined time points throughout 500 my of evolution (Fig. 1). In most species, there is an embryonic α-like gene(s) (α), adult gene(s) (α), and theta gene(s) (θ). In addition, we found that, in most species, there is a functional homologue of the previously characterized avian αD gene (20), lying between the θ and α genes; in man, this gene was previously annotated as a pseudogene (iα2; ref. 21). The structure and evolution of the α-like genes are summarized in Figs. 1, 2, and 5 (see also Table 2, which is published as supporting information on the PNAS web site).

We have previously shown that the human α cluster lies in a GC-rich, repeat-rich, gene dense region of the genome close to the telomere of the short arm of chromosome 16, associated with internal, direct repeats of telomeric DNA (TTAGGG)n characteristic of such regions (22). The α-globin cluster in the rabbit (24) and horse (25) are also known to lie close to telomeric regions. Here, we found telomeric repeats upstream of the α cluster in 8 of 16 species analyzed, suggesting that, in many species, it is located or originated close to a telomere; a clear exception occurs in the mouse (Fig. 1). Between compared species, the extent of each black bar indicates the amount of DNA sequenced in each species.

The Evolution of CpG Islands and Their Role as Cis Elements. Within this region, CpG islands are among the simplest of cis elements to identify. They are short, unmethylated, GC-rich, CpG-dense regions of DNA usually located at the 5′ ends of genes and, potentially, they mark the position of a large proportion of promoters in mammalian genomes. Their role, if any, in regulating normal gene expression is not clear (29). We have previously characterized the CpG islands associated with the tissue-specific human α-globin genes and the widely expressed genes that flank them (10) (A–K, Fig. 6, which is published as supporting information on the PNAS web site). The loci are aligned on the highly conserved sixth exon of the C16orf135 gene, which is indicated by an asterisk.
a general lack of CpG suppression in the α cluster of this species (Fig. 5). A major difference was observed in rodents in which, as noted in ref. 31, some islands (including H9251 globin) were substantially eroded or absent, whereas others (e.g., C and E) remained readily detectable. Whatever the mechanism by which they are lost, it appears that CpG islands are certainly not essential for gene regulation and, although often a valuable marker, may not provide a consistent guide to the presence of regulatory elements in all species.

The Identification of Multispecies Conserved Sequences. To identify MCSs, we initially used a global alignment (AVID visualized with VISTA, see Fig. 3), identifying and noting all conserved, noncoding regions that showed a minimum of 50% conservation in multiple mammalian species. VISTA highlights, in pink, regions with >75% conservation in a 100-bp window. Conservation >75% is colored pink, and annotated exons are shaded gray. At the top, shown are three classes of MCS: MCS-R1–3, MCS-S2, and MCS-E2. The position of the erythroid DHSs, HS-48, HS-40 and HS-33, are indicated by red arrows above. The direction of transcription (black arrow) and positions of exons (blue boxes) of the C16orf35 gene are shown below the VISTA plot.
region of 238 kb. These MCSs were identified entirely independently of any prior knowledge of previously characterized regulatory elements in this area. For each MCS, we scored the length of conservation, noted all of the species in which conservation could be detected, and calculated the degree of conservation (Table 1).

To compare this approach with current algorithms, we searched for MCSs by using WEBMCS (9) set at the default value (95%), which assumes that 5% of the genome is under purifying selection. When all known exons were excluded, WEBMCS identified 13 individual MCSs, including 12 of the 24 MCSs described here. When set at 90%, WEBMCS found 138 elements, including 23 of the 24 MCS elements identified here. To search for MCSs in an another independent manner, we also ran GUMBY on MULTILAGAN alignments of closely related eutheriens in the region of conserved synteny and the adjacent region containing Luc7L. At the default P value threshold of 0.5, GUMBY identified 20 of the 24 MCSs defined here and 19 additional regions. At a stricter threshold of 0.1, GUMBY identified 16 of the 24 MCS and 10 other predictions. These findings are discussed in detail elsewhere (S. Prabhakar, unpublished data).

**Classification of MCSs.** Together with previously described mapping of CpG islands (10, 22), localization of DHSs (10, 11), identification of ESTs, and annotation of genes (22), the 24 MCSs identified here could be subclassified into five groups. These groups are promoters (MCS-Ps), other transcriptional regulatory elements (MCS-Rs), splicing elements (MCS-Ss) previously unidentified exons (MCS-Es) and unknown elements (MCS-Us) (Fig. 2 and Table 1).

We found 11 MCSs associated with known promoters. Four (MCS-P1, MCS-P2, MCS-P6, and MCS-P11; Fig. 2) were found within CpG islands associated with the constitutively expressed DHSs of widely expressed genes. MCS-P3–MCS-P5 were found associated with a DHS marking an alternative promoter of the gene encoding methyladene DNA glycosylase (gene 6 in Fig. 2). Conserved elements MCS-P7–MCS-P10 were found in the promoters and CpG islands (when present) of the φ, αD, α, and θ globin genes, respectively. Each of these genes is associated with erythroid-specific (φ, αD, and α) or constitutive (θ) DHSs. Of the globin gene promoters, the embryonic (φ) globin gene was the most highly conserved (64% conservation over 367 bp between man and mouse and detectable in all mammals).

A second group of MCSs (MCS-R1–MCS-R4 in Fig. 2 and Table 1) corresponded to known and presumed cis-acting regulatory elements associated with erythroid-specific DHSs. MCS-R1 identified a previously uncharacterized erythroid-specific DHS (ref. 16 and M.D.G., E.A., and J.R.H., unpublished data). Most of these MCSs were conserved, to variable extents, in all mammals (68–100%, Table 1) although MCS-R3 was not detectable in cows. MCS-R2 was the most highly conserved regulatory element, based on its 69% conservation between mouse and human over a region of 256 bp, and the fact that it could be detected in all mammals. This finding is consistent with its previously established role as the human α-globin major regulatory element (αMRE). None of these elements was readily detectable in birds (310 my) or fish (415–485 my), although the same repertoire of TFBSs found in MCS-R2 were also found in the orthologous region in these vertebrate groups (28).

A third group includes MCS-S1 and MCS-S2. Initially, the role of these elements was not clear because they were not represented as ESTs and did not associate with any known DHSs (Table 1). However, we noted that MCS-S2 lies close to an alternative coding exon (MCS-E2, see below) within the gene C16orf35 (gene 7 in Figs. 2 and 3) but does not itself contain an ORF and has no EST

<table>
<thead>
<tr>
<th>Table 1. Conservation of MCSs in 22 species</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>---</td>
</tr>
<tr>
<td>Length (bp)</td>
</tr>
<tr>
<td>HS Human</td>
</tr>
<tr>
<td>HS Mouse</td>
</tr>
<tr>
<td>EST</td>
</tr>
</tbody>
</table>

The conservation of the described MCSs is shown for each species. MCSs are separated into five subclasses (see text). The length, names of the DHSs or promoters with which the MCS are associated in human and mouse, and whether it is transcribed are shown in rows 2–5, respectively. For each species, MCSs are labelled as follows: +, present; N, not detectable; (-), gene not orthologous; (+), similar TFBS but not homologous; NS, no sequence. Values for % conservation in human, mouse, and rat (HMR) and proportion of bases perfectly conserved in all species are given. The asterisk in MCS-P7 indicates this figure was derived from an alignment without hedgehog and opossum due to evolutionary shuffling of the conserved motifs in these species. + indicates in which percentile webMCS detected the described MCS and, at right, the number of additional MCSs scored. Coordinates for start and stop of each MCS are given relative to the telomere of human chromosome 16.
Fig. 4. Examples of conserved TFBs. The example is from MCS-R2. (a) A fully conserved GATA binding site followed by two conserved Maf recognition elements (MAREs) that form the core of this MCS. (b) A conserved GATA binding site in MCS-R2 that is lost in rodents and hedgehogs.

database matches (UCSC/Ensembl genome browsers and NCBI BLAST). These data, combined with the relatively high degree of conservation in this element (MCS-S2 is 85% conserved over 158 bp between human and mouse and is detectable in most mammals, opossum, and chickens; Table 1) is reminiscent of conserved elements thought to be associated with RNA processing, although it does not fulfil the strict criteria for an ultraconserved element (32). Support for such a role in RNA splicing came from multispecies analysis. In cows, pigs, and hedgehogs, the only evolutionary lines in which the neighboring coding exon MCS-E2 is not conserved (Table 1), MCS-S2 has also undergone rapid sequence divergence, suggesting that conservation of MCS-S2 is required to act as a signal for the alternative splicing of MCS-E2. In support of this hypothesis, a phylogenetic footprint of MCS-S2 revealed two highly conserved (100% from human to chicken) sequences (AG-CATG), previously shown to be functional components of intronic control regions (ICRs), which regulate alternative splicing (33, 34). Similar conserved ICRs were also found in MCS-S1, which lies close to another known alternatively spliced exon in the rhomboid-like gene (gene 5, Fig. 2).

Four conserved elements (MCS-E1–MCS-E4) were present within spliced ESTs and are thought to represent exons. MCS-E1 is an alternative coding exon of the rhomboid-like gene (gene 5 in Fig. 2), and MCS-E2 is an alternative coding exon of C16orf35 (gene 7 in Fig. 2). MCS-E3 is an alternative coding exon of the LUC7L gene (gene 16 in Fig. 2), which encodes a component of the U1snr complex (35). MCS-E4 is a large, highly conserved element (1,600 bp) lying in the first intron of LUC7L. Part of MCS-E4 is incorporated as a noncoding exon, although its inclusion disrupts the normal reading frame of LUC7L. A 204-bp segment of MCS-E4 fulfills the criteria for an ultraconserved element (32) with only one base pair difference between the human, mouse, and rat.

We could not assign any known function to the remaining three elements MCS-U1–MCS-U3 (Fig. 2) even though they appear to be highly conserved (Table 1).

Erythroid MCSs Contain Clusters of Conserved TFBs. Because expression of the α-globin cluster has been particularly well characterized, we analyzed the association of MCSs with erythroid-specific DHSs in depth. In particular, we optimized alignments of each MCS corresponding to the erythroid-specific promoters (MCS-P7–MCS-P10) and regulatory elements (MCS-R1–MCS-R4) and analyzed them by using the TRANSFAC database (version 6.0), both directly and utilizing TREs and R-VISTA, looking for proteins that might bind these conserved elements. Many of the phylogenetically conserved motifs corresponded to known sites of functional importance that have been previously shown to bind transcription factors in vitro and in vivo (16, 36–38). In particular, we found that six (MCS-R1–MCS-R4 and MCS-P7 and MCS-P8) of the eight MCSs associated with erythroid-specific DHSs contain at least one very highly conserved GATA binding site (WGATAR). Such sites are characteristic of elements regulating genes that are switched on or off during erythropoiesis (39). MCS-R2 (αMRE) also contained two highly conserved Maf regulatory elements (40). MCS-P9 (α-globin promoter) contained conserved CCAAT and ATA motifs but no conserved GATA binding sites. Some elements (MCS-P7, MCS-P8, and MCS-P10) also contained highly conserved GC-rich elements, and some MCSs also included highly conserved elements that do not obviously correspond to the binding sites for any known transcription factors.

In many regulatory MCSs, individual TFBs have been completely conserved across all species (e.g., GATA binding site in MCS-R2, Fig. 4a). Others are fully conserved in most species but have mutated one or more critical residues that would alter binding in one or two species (e.g., GATA binding site in MCS-R2, Fig. 5b). In some MCSs, we noted that the disappearance of a conserved GATA site in one position was associated with the appearance of a new GATA site elsewhere in the element (e.g., in the dog in MCS-R4). Although the function of these new sites has not yet been tested, these observations are consistent with the concept of transcription factor “turnover” and coevolution (41), in which a regulatory element conserves overall function, albeit by using different sequence motifs.

Spacing Between Conserved TFBs in MCSs. From these and other data (16, 42), it seems likely that MCS-R1–MCS-R4 and MCS-P7 and MCS-P8 act as binding sites for multiple interacting proteins regulating gene expression and/or chromosome function. If such proteins bind and/or interact cooperatively, the spacing between binding sites may also have been conserved. The distances between the most consistently conserved binding sites (Fig. 7a, which is published as supporting information on the PNAS web site) in each MCS are plotted in Fig. 7b. In most elements, the distances between such sites are variable. However, in MCS-R2 (αMRE), the spacing and sequences between the fully conserved GATA binding site and two Maf recognition elements (Fig. 4 and intervals IIb and Iic in Fig. 7a) which form the core of this MCS, are highly conserved in 18 species, whereas spacing of the flanking GATA sites (intervals Ha and Hf in Fig. 7a) are not (Fig. 7b). This result suggests that not only the sequence, but also the spacing and orientation of TFBs on the DNA helix, are highly conserved in some MCSs, which correspond to key regulatory elements.

Discussion

This study demonstrates that a combination of CpG analysis, mapping of DHSs and evaluation of MCSs, as defined here, can identify and subclassify most, if not all, critical, noncoding, cis-acting sequences. The additional question we have asked here is to what extent MCS analysis alone could achieve this result.

The multispecies analysis described here was highly sensitive and accurate for independently identifying previously characterized cis-acting elements within and surrounding the human α-globin
These elements included 12 of 20 fully mapped constitutive and erythroid-specific DHSs. Further close inspection of the sequences underlying the remaining eight DHS showed that although three are associated with CpG islands, none contains an MCS. It is of interest that whereas most DHS can be detected in humans and mice (16, 45) four of the seven DHS analyzed were not found in mice. Furthermore, in contrast to the 17 DHS associated with MCSs and/or CpG islands, there is no observational or experimental data to show that the remaining 3 DHSs (+37, −8, and −55) are of functional importance.

It is of interest that by using the standard parameters and same data set, WEBMCS identified 13 conserved elements, including 12 of the 24 MCSs described here and one additional element that failed our criteria. Several of the 24 MCS elements were missed, including some of known functional significance. Most elements were found by using a less stringent WEBMCS search (Table 1), but this search also identified 20 of the 24 MCSs and 19 additional candidates. Therefore, it appears that judicious use of current algorithms for identifying MCSs can detect many cis elements with good sensitivity and without an overwhelming number of false positives.

Independently of WEBMCS and GUMBY, the simple but stringent bioinformatic approach described here identified most of the key cis elements in a previously well characterized segment of the genome. Whether these criteria would identify such elements with similar efficiency in other areas of the genome remains to be tested. Classification of MCSs clearly requires integration with other bioinformatic analyses and experimental data, particularly identification of ESTs and DHSs as described here.

Sequences of the chimpanzee, cow, and cat were from ENCODE. We thank The Oxford Computational Biology Research Group for support and Prof. W. Wood, Dr. R. Gibbons, Prof. R. Patient, and Dr. M. de Bruijn for helpful comments. This work was funded by the Medical Research Council (U.K.) and partly supported by National Heart, Lung, and Blood Institute Programs for Genomic Applications Grant HL66728 (to E.R. and J.-F.C.).